https://doi.org/10.46938/tv.2022.552

////// studie / article //

INTERNET JAKO PRAMEN VÝZKUMU: PŘÍSTUP K ARCHIVOVANÝM WEBOVÝM ZDROJŮM A MOŽNOSTI JEJICH ZPRACOVÁNÍ

Abstrakt: *Internet se stal přirozenou komunikační platformou soudobé společnosti. Webové archivy, které začaly vznikat v 90. letech 20. století s cílem zachytit a uchovat proměnlivý webový obsah, se tak staly klíčovými prameny pro výzkum nedávné minulosti. Analyzování jejich dat komplikují například nedostatečné kompetence badatelů, nutnost vybavení výkonnými výpočetními zdroji nebo legislativa. Jednou z cest, jak vyjít vstříc potřebám uživatelů, je vývoj nástrojů a výzkumných rozhraní, které umožňují práci s daty bez nutnosti technologických znalostí pokročilé extrakce a otevírají je tak k využití badatelům. Studie řeší problematiku zpřístupnění archivních webových dat, přibližuje snahy o formulování teoretického a metodologického rámce a navrhuje design pro přístup a pro další zpracování dat, který je aplikován v unikátním výzkumném rozhraní pro vytěžování velkých dat z webových archivů s využitím pokročilých postupů strojového zpracování pro generování a kategorizaci textových výstupů.*

Klíčová slova: *archivace webu; Webarchiv; vytěžování dat; datová analýza; výzkumná rozhraní; Hadoop*

ZDENKO VOZÁR
Národní knihovna ČR
Klementinum 190, 110 00 Praha 1,
Ústav religionistiky FF MU
Arna Nováka 1, 602 00 Brno
email / zdenko.vozar@gmail.com
0000-0001-8414-7939

Internet as a Source of Research: Access to Archived Web Resources and Possibilities of Their Processing

Abstract: *The Internet has become a natural communication platform for modern society. Web archives, which began in the 1990s to capture and preserve changing web content, have thus become key sources for research in the recent past. The analysis of their data is complicated by, for example, insufficient competencies of researchers, the need for computing resources or legislation. One way to meet the needs of users is to develop tools and research interfaces that allow to work with data without the need for technological knowledge of advanced extraction and thus open them to researchers. The study addresses the issue of access to archival web data, approaches efforts to formulate a theoretical and methodological framework and proposes a design for access and further data processing. This design is applied in a unique research interface for extracting large data from web archives using advanced machine learning to generate and categorization of text outputs.*

Keywords: *web archiving; Webarchiv; data mining; data analysis; research interfaces; Hadoop*

MARIE HAŠKOVCOVÁ
Národní knihovna ČR
Klementinum 190, 110 00 Praha 1
email / marie.haskovcova@nkp.cz

ANDREA PROKOPOVÁ
Národní knihovna ČR
Klementinum 190, 110 00 Praha 1
email / ap.prokopova@gmail.com

Článek byl realizován v rámci výstupů projektu „Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů" č. DG18P02OVV016 financovaném z programu na podporu aplikovaného výzkumu Ministerstva kultury ČR NAKI II.

1. Úvod

V posledních dekádách došlo k zásadnímu rozšíření pramenné základny pro studium člověka a lidských dějin. Fenomén internetu vytvořil z uzavřené sítě úzkých komunit specialistů globální fenomén písemné a multimediální veřejné komunikace. Technologické změny a významné veřejné sociální platformy, například Facebook, YouTube, Twitter, blogy a další kanály ulehčují možnost instantní komunikace, masivní sebeprezentace (tvorby stránek, profilů, skupin) a zároveň bezprecedentní dynamickou tvorbu nebo sdílení obsahu. Co se týká informačního obsahu, významně narostly možnosti jeho globálního zpřístupňování (např. Google indexace) nebo konzultování a řízení kvality veřejných znalostních bází (např. Wikipedia). Internet je tak dnes bezesporu nedílnou součástí pracovního i osobního života a obsahuje velké množství veřejně dostupných aktuálních a rychle zastarávajících informací. To se potvrdilo i během masového přesunu aktivit lidí do virtuálního prostředí v prvních měsících pandemie onemocnění koronavirem COVID-19. Ať už mluvíme o webových stránkách vládních institucí, firem nebo o reklamách a osobních blozích, všechny tyto zdroje mohou mít pro výzkumníky v budoucnu nevyčíslitelnou hodnotu jako obraz soudobé společnosti. Přesun části života člověka a jeho sociálních skupin do online prostředí a s tím související zrychlování informační výměny však pro badatele otevírá širokou sadu otázek, jakým způsobem pracovat s rychle vznikajícími a často prchavými prameny v řádech miliard jednotek, kterými je nutné se koncepčně, ale i technicky zaobírat.

Na rozdíl od tradičních pramenů, u kterých může být problémem jejich malé množství, nízká kvalita, nebo nedostatek informací o jejich vzniku, tvůrcích a organickém kontextu, u informací sbíraných z internetu je problémem obrovské množství různě relevantních dat, velice obtížně zpracovatelných tradičními postupy historické a literární textové kritiky. Jde však o největší zdroj písemných a obrazových materiálů v dějinách lidstva pocházející od všech vrstev společnosti, a proto je žádoucí tento problém vnímat jako příležitost. Jak uvádějí Niels Brügger a Ian Milligan,[1] web je současně historický zdroj i předmět studia. Podle jejich pojetí webová historie (*web history*) odkazuje nejen k historii webových stránek nebo internetu, ale zahrnuje také metodické úsilí historiků, kteří používají webové archivy ke své práci. Osvobození od fyzického formátu rovněž vede k novým otázkám ohledně postupů verifikace autenticity. Další výraznou charakteris-

[1] Ian Milligan, „You Shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure," WARCnet Papers, 2020, 1–17.

tikou internetových pramenů je jejich krátký poločas rozpadu. Specificky u webových stránek výzkumy uvádějí, že jejich životnost je okolo 100 dní. Následně jsou stránky upraveny, aktualizovány nebo přesunuty, a mohou tak být touto manipulací znehodnoceny.[2] Zcela zásadní se tak ukazuje snaha o konzervaci informací z prostředí internetu prostřednictvím plánovaného získávání objektů a uchovávání jejich kontextu. Předpokládáme, že tato data se brzy stanou významným pramenem k pochopení současné historie, který nebude možné přehlížet. Aby se tak ale mohlo stát, je nutné data archivovat do repozitáře, zajistit autenticitu jejich kopií a následně je zpřístupnit historikům a celkově badatelům v oblasti humanitních věd a digital humanities.

O archivaci, verzování, dlouhodobé uložení a zpřístupňování těchto informací se pokoušejí národní knihovny a další paměťové instituce a iniciativy po celém světě. Pro digitalizované knihy nebo jiné fyzické artefakty se jim podařilo vytvořit repozitáře, které jsou již v dnešní době relativně lehce dosažitelné a filtrovatelné (např. statisíce monografií a periodik na jednom místě dostupné od počítače). Archivní webová data obsahují zásadní informace k dějinám, sociálním a kulturním pohybům 21. století, podobně jako rozsáhlé a již dostupné digitalizované prameny k 20. století a hlubší minulosti. Otázkou však je, nakolik bude u těchto dat metodicky zachována jejich autenticita, integrita a zda budou dostupná na základě požadavků odborné, ale i laické veřejnosti. Pokud budeme pokračovat v naznačené analogii, tak zatímco k fyzickému vydání digitalizované knihy se lze vrátit, informace z internetu je efemérní a její zánik může být téměř okamžitý. Na jedné straně je nutné vytvořit nový kritický aparát pro vypořádání se s novým typem pramenů a zároveň pochopit, jaké jsou lidské, technické a právní podmínky jejich sběru. Ty mají zásadní dopady na interpretační úsilí archivních dat. Na druhé straně je potřeba vybudovat badatelsky srozumitelné, avšak technicky poměrně složité rozhraní, které se může stát pevným bodem kvantitativních i kvalitativních metodik vedených nad takto získanými daty a metadaty.

Z hlediska kompetencí vědců je jednou z možností, jak s tímto typem pramenů pracovat, významné rozšíření jejich technických schopností, nebo formování širokých multidisciplinárních týmů, které budou zahrnovat i IT odborníky. Tento postup je však pro šířeji koncipovaný výzkum webových pramenů velkou výzvou, nebo dokonce bariérou například pro tradičně orientované humanitní vědce. Metodicky pracují spíše individuálně než týmově a kladou důraz na kvalitativní postupy – na rozdíl od technicky za-

[2] Ester Shein, „Preserving the Internet," *Communications of the ACM* 59, no. 1 (2016): 26–28.

měřených kolegů, nebo i badatelů v oblasti sociálních věd. Zejména odborně vedená analýza nad digitálními daty může být převážně kvantitativního typu, ale do budoucna může nabývat více kvalitativního charakteru. Pro její provádění je však nutné překonat více nejen technických, ale i metodologických, kompetenčních i právních bariér. Domníváme se, že v tomto ohledu bude nutné vyjít vstříc různým uživatelským skupinám, jejich badatelským požadavkům a jejich organizačnímu a metodickému zázemí.

Je nutné si však uvědomit i značné kapacitní nároky na práci s tímto typem dat. Například pro textovou extrakci je nutné použít relativně velkých výpočetních prostředků nad objemnými „surovými daty". Významné snížení technické bariéry pro základní práci s datovým obsahem nastává díky integraci prvků workflow nebo poskytnutím výpočetních prostředků nad zpracováním velkých dat, které by na pracovním počítači nebyly myslitelné, nebo v cloudu finančně dostupné. Tyto potřeby se snaží naplňovat v dnešní době zejména projekt Archives Unleashed Cloud[3] (dále AUC) prostřednictvím různě zaměřených pracovních prostředí pro vědce s různou úrovní technických kompetencí.[4] Podmínkou je přitom nutnost vkládání vlastních dat. Podobné projekty vedou k dramatickému zvýšení dostupnosti pramenů archivovaného webového obsahu, ať už pro etapu identifikace pramenné základny výzkumníkem, tedy vymezení datasetu, export „surových dat", nebo jejich předzpracování a operativní korpusový management.

Důležité bude proto koncepčně umožnit práci s rozsáhlými datovými soubory, stejně jako růst individuálních projektů, a to pomocí snížení kompetenční bariéry a možnosti definice vlastních specifických datasetů. Tento postup vede ke snaze o designování procesů přístupů pro průzkum konkrétních webových archivů, které zásadně zjednoduší orientaci v jejich obsahu a extrakci materiálu. V rámci potřeb Českého webového archivu Národní knihovny ČR (Webarchivu),[5] který se zabývá archivací českého internetu, vznikl v roce 2018 projekt Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů,[6] který se pokouší vytvořit rozhraní pro průzkum, manipulaci a export textového obsahu a kontextu archivních dat pro různé typy badatelů, aby archiv dále mohl být analyzován. Na jedné straně umožní zapojit Webarchiv jako zdroj dat do velkých

[3] The Archives Unleashed Project, https://archivesunleashed.org.
[4] Milligan, „You Shouldn't Need to be a Web Historian to Use Web Archives," 1–17.
[5] Národní knihovna České republiky, „Webarchiv," https://www.webarchiv.cz.
[6] Projekt evidovaný pod číslem DG18P02OVV016 je financován ze zdrojů dotačního mechanismu Ministerstva kultury NAKI II. Doba realizace projektu je pětiletá, od roku 2018 do konce roku 2022.

výzkumných infrastruktur a datových architektur a zároveň i chápat data v kontextu podmínek jejich vzniku,[7] datového setu, sbírkové politiky a institucionálního ukotvení. Předpokládáme, že designování prostředí pro práci s internetovým obsahem jako historickým pramenem může mít zpětně dopad na metodiku práce s tímto typem pramenů, podobně jako archivy a archivnictví měly dopad na formování klasické historie. Proto chceme nabídnout co nejširší možnosti definice vlastních datasetů a čerpání a využívání kontextů vzniku dat.

2. Webové archivy jako instituce a jejich datová základna

Rostoucí význam webové archivace je neoddiskutovatelným trendem informační společnosti. Jednou z prvních institucí, která se začala archivaci internetu věnovat, je americký Internet Archive,[8] který v roce 1996 založil Brewster Kahle. Tato unikátní soukromá digitální knihovna obsahuje nejen archivní kopie webových stránek z celého světa, ale i knihy převedené do digitálních formátů, audio a video záznamy, počítačové hry nebo softwarové programy. Vznikla s vizí vytvořit veřejnou knihovnu dostupnou každému, kdo má přístup na internet a intenzivně ji využívají jak výzkumníci, tak široká veřejnost.[9] V roce 2003 se 12 institucí v čele s Internet Archive rozhodlo spojit své síly a vytvořit mezinárodní platformu webových archivů International Internet Preservation Consortium[10] s cílem koordinovat úsilí v oblasti zachování webového obsahu. Konsorcium se zasazuje o sdílení zkušeností, ať už se jedná o sklízení, uchovávání a zpřístupňování webu, o legislativní nebo etické aspekty. Má normativní funkci, usiluje o formulování tzv. *best practices*, vytváření nástrojů a standardů, které umožňují společný vývoj a spolupráci. Definovalo klíčový standard – kontejnerový formát pro uložení dat, z nějž vycházejí všechny snahy o jejich badatelské zhodnocení.[11] Konsorcium v současnosti tvoří více než 50 institucí. V rámci České republiky archivaci webu zajišťuje webový archiv Národní knihovny ČR. Od roku 2000

[7] Například doména, na kterou je odkazováno, je archivována častěji, dochází k personalizaci obsahu, sklízí se pouze veřejný obsah, dostupné nástroje mají problém se sklízením dynamického obsahu atp.
[8] Internet Archive, https://archive.org.
[9] Miguel Costa, Daniel Gomes, and Mário J. Silva, „The Evolution of Web Archiving," *International Journal on Digital Libraries* 18, no. 3 (2017): 191–205.
[10] IIPC, „International Internet Preservation Consortium," https://netpreserve.org.
[11] Jedná se o balíček WARC, dostupný jako ISO 28500:2009. „Information and Documentation – WARC File Format," https://www.iso.org/standard/44717.html.

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.[12] Jak je zřejmé z výčtu členů konsorcia IIPC, k národním institucím odpovědným za archivaci webu patří často národní knihovny, podobně jako v případě českého webového archivu. Může se ale také jednat o výzkumné instituce, jako v případě portugalského archivu,[13] který je propojen s portugalskou národní sítí pro výzkum a vzdělávání, nebo o soukromé subjekty. Protože mnoho webových archivů funguje na národní úrovni, pracují s konceptem národního webu. Pohybují se v různých právních rámcích a řada z nich má oprávnění získávat obsah vytvořený ve svých zemích svými občany. Za jasný identifikátor je považována národní doména, avšak národní legislativy, jako je například francouzská, dánská nebo česká, neomezují národní web pouze na ni (viz definice bohemikálního dokumentu v následující kapitole). Vymezení národního webu a zacílení akviziční strategie je klíčové pro formulování výzkumných otázek i interpretace bádání.

Správci webových archivů si uvědomují, že kromě primárních funkcí, jako je sběr, ukládání, ochrana a zpřístupňování archivních dat je pro jejich budoucí využití důležitá i jejich kontextualizace, schopnost datům a jejich hierarchii porozumět, jak zmiňují například Valérie Schafer a Jane Winters.[14] Zamýšlejí se nad výzvami udržitelnosti, odpovědnosti, etické a veřejné angažovanosti i nad požadavkem na rozšiřující se spektrum zúčastněných stran. Vznikají snahy o formulování teoretického a metodologického rámce nebo návrhů workflow, jak s daty pracovat. Snažil se o to například projekt britského webového archivu Buddah[15] zaměřený na oblast umění a humanitních věd na britské doméně v období od 1996 do 2013. Na základě tohoto projektu vzniklo SHINE – interaktivní rozhraní pro dotazování se na frekvence jmenných entit, korpusové výskyty a analýzu trendů pomocí fasetového vyhledávání.[16] Z novějších jmenujme například belgický projekt

[12] Jaroslav Kvasnica et al., „Analýza českého webového archivu: Provenience, autenticita a technické parametry," *ProInflow* 11, č. 1 (2019): 3–21.
[13] Arquivo.pt, „Information about the Arquivo.pt service," https://sobre.arquivo.pt/en.
[14] Valérie Schafer and Jane Winters, „The Values of Web Archives," *International Journal of Digital Humanities* 2 (2021): 129–44.
[15] University of London, „Big UK Domain Data for the Arts and Humanities," https://buddah.projects.history.ac.uk.
[16] UK Web Archive, „SHINE," https://www.webarchive.org.uk/shine.

Besocial,[17] který se věnuje metodologii archivace sociálních médií, nebo vědeckou platformu WARCNet[18] podporující výzkum digitálního kulturního dědictví národních webových archivů. Diskutovány jsou i možnosti, jak badatelům zprostředkovat data napříč archivy. S ohledem na restrikce v národních legislativách se jeví jako jedno z možných řešení vytvoření nadnárodního úložiště metadat, které by umožnilo vědcům pracovat s daty napříč archivy.[19] Webové archivy mají velký potenciál pro budoucí výzkumy. Aby bylo možné jejich data využít, je však potřeba hledat řešení, jak lze s daty pracovat a zpřístupňovat je jako datové základny pro studium současných dějin, jazyka, kultury i jako technické muzeum – památník internetu a jeho technologií.

3. Otázka reprezentativnosti a dostupnosti dat webových archivů

Webové archivy mohou sloužit jako bohatý zdroj informací pro vědecké záměry, archivní webová data však mají svá specifika. Dle Juliena Masanese webová stránka sama o sobě nikdy nedává plně smysl, je vmíchaná do organického kontextu – sítě dokumentů, které mezi sebou referují a až společně tvoří kontext.[20] Zásadním aspektem archivovaného digitálního obsahu je tedy jeho autenticita. Lze ji charakterizovat tak, že dokument je tím, za co se vydává, nebyl zfalšován nebo porušen.[21]

3.1 Povaha reprezentativnosti dat

U webových dat je riziko porušení autenticity, tedy uchování zdroje a jeho pravosti tak, jak se zobrazuje uživateli v konkrétním momentě, vysoké už kvůli povaze vzniku informace, technologii jejího zachycení, jejímu dalšímu dlouhodobému uchovávání a neporušené prezentaci i po letech. Webové stránky jsou dnes generovány dynamicky a kontextově a obsahují velké množství prvků. S ohledem na informace o charakteru samotného uživatele se jednotlivé části mohou měnit. Poskytovatel může přesunout obsah na jinou doménu, může docházet k personalizaci obsahu dle polohy

[17] The Royal Library of Belgium, „Besocial," https://www.kbr.be/en/projects/besocial.
[18] Aarhaus University, School of Communication and Culture, „About WARCnet: Web ARChive Studies Network Researching Web Domains and Events," https://cc.au.dk/en/warcnet/about.
[19] Schafer and Winters, „Values of Web Archives."
[20] Julien Masanes, „Web Archiving Methods and Approaches: A Comparative Study," *Library Trends* 54, no. 1 (2005): 72–90.
[21] Ladislav Cubr, *Autenticita a digitální informace* (Praha: Univerzita Karlova v Praze, 2017).

nebo přístupové IP adresy. Specifická skupina problémů souvisí s etikou výběru obsahu a cílů sklizně a nastavení limitujících algoritmů pro sběr. Tato témata technicky a kurátorsky souvisejí také limitním obsahem, jako je uchovávání a potenciální pozdější šíření závadného obsahu narušujícího soukromí a bezpečnost uživatelů (např. technického – malware nebo informačního – dezinformací). Hodnocení pravdivosti obsahu zdroje na rozdíl od jeho pravosti kurátorům webových archivů nepřísluší. Dalším důležitým aspektem je provenience dat a metadat. V obecné rovině se jedná o původ a zachování celistvosti informací, přičemž dostatek relevantních metadat zásadním způsobem zvyšuje reprezentativní hodnotu zkoumaných datasetů a pochopení jejich limitů.[22] V současné době nejsou výjimkou datasety bez dochovaného metadatového a věcného popisu. Pro konkrétní badatelské využití tak může absence těchto informací představovat významnou překážku pro další analýzy. Je důležité, aby badatelé měli k dispozici jak data, tak i maximum kontextuálních metadat.

Pro reprezentativnost a dostupnost dat je klíčový způsob akvizice. Z technického hlediska se jedná o automatické získávání obsahu z internetových serverů – tzv. sklízení (*harvesting*). Domněnka, že je web statický, neplatí už od doby jeho vzniku. S dnešními technologiemi tvorby a správy webu je změna obsahu daleko snazší a jak již bylo zmíněno, obsah je často přizpůsobován prohlížiteli. K nežádoucím nebo těžko ovlivnitelným změnám sklízeného obsahu může dojít již během různých fází přístupu sklízeče – ještě před samotným zachycením dat na základě např. profilování dle regionů a dalších informací o uživateli. To představuje problém nejen při prohlížení, ale i při sklízení, kdy jsou s ohledem na politiku postupného vytěžování „sbírány" (harvestovány) elementy stránky v časově postupném sledu. Jakým způsobem je možné zachytit exaktně informaci ve všech jejích manifestacích a podobách? Kvůli technickým omezením většina webových archivů postupuje tzv. minimalistickou metodou standardního uživatele a standardního prohlížeče, tak aby získala alespoň jednu, co nejúplnější verzi. Právě zde se nachází určité omezení autenticity. Stejně tak je nutné počítat s jistou mírou neurčitosti, respektive neúplnosti všech verzí informace tam, kde se informace skládá dynamicky vzhledem k pozici pozorujícího. Hledat zde úplnou autenticitu je iluzorní, proto je nutné vnímat zachycený celek vždy jako neúplnou informaci, odpovídající pozici pozorovatele a politikám nejen vydavatelů. Tato neurčitost se projevuje zejména v kontextové

[22] Niels Brügger and Ralpf Schroeder, eds., *The Web as History* (London: UCL Press, 2017).

reklamě, ale také v obsahu, který odpovídá komunikačním schématům many-to-many, kde garance jedné verze ztrácí smysl.

Rizika změny informace, i když technologicky minimalizovaná, jsou rovněž přítomná při ukládání dat, jejich zpracování, dlouhodobém uchovávání a jejich prezentaci. Jde ovšem již o ovlivnitelná a zmírnitelná in-house rizika: zejména jde o konflikty se síťovým tokem sběru dat proti limitované šířce pásma celé organizace zpožďující sběr, případně o změny při migraci dat hierarchického úložiště, změny hardwaru a technologií, nebo použitou metodiku pro obnovu ztracených dat ze záloh. I když jsou data bezpečně zachycena a správně uložena, pro uživatele může problém nastat při změně prezentační technologie s nižší mírou zpětné kompatibility, nebo v případě, že nebyly všechny části stránky plně zachyceny a chybí např. důležitá knihovna třetí strany, která byla uložená mimo archivovanou doménu. Webové archivy proto musí vytvářet a uchovávat dokumentaci, která popisuje nejen způsob vzniku archivních kopií, ze které budou zřejmé jejich limity a změny v čase, nýbrž i koncepce jejich ukládání a uchování, včetně metadatových standardů doporučujících povinné a volitelné informace o jejich entitách: sklizních, datových kontejnerech a lozích každé sklizně.[23] Tato dokumentace by měla také zaznamenávat způsoby prezentace a emulace prohlížení. Samotné prohlížení je velice omezeno a k emulaci používá verze softwaru Wayback Machine, který umožňuje vyhledávání jen při znalosti URL adresy stránky,[24] což je nevyhovující pro kvantitativní i kvalitativní analýzu stránek, pokud badatel nezná weby, v nichž by mohl zkoumat určité téma.

3.2 Strategie budování českého webového archivu

Každý archiv, i ten webový, by měl být budován s určitou vizí, a tato vize přímo ovlivňuje výslednou podobu archivních dat. Cílem webové archivace je výběr, uchovávání a zpřístupnění dat uživatelům, čili budování komplexního fondu digitálních zdrojů.[25] Webarchiv se v tomto případě řídí dokumentem Strategie budování sbírky Webarchivu (*collection policy*),[26] který hlouběji popisuje hlavní cíle Webarchivu, typy sklizní, kritéria výběru dokumentů i definici bohemikálního webu. Představuje pravidla, zásady a cíle,

[23] Jaroslav Kvasnica et al. *Metodika pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu* (Praha: Národní knihovna ČR, 2020).
[24] V současnosti se postupně přechází na software Pywayback.
[25] Kvasnica et al., „Analýza českého webového archivu," 3–21.
[26] Kvasnica et al., „Strategie budování sbírky Webarchivu. Aktualizované znění," Národní knihovna České republiky, 2019.

podle kterých je archiv budován. A tato pravidla přímo ovlivňují podobu a vlastnosti sklizených dat. Činnost Webarchivu se opírá o knihovní licenci Autorského zákona,[27] která umožňuje vytváření sbírek webového obsahu pro konzervační a archivační účely.[28] K akvizici zdrojů využívá tři různé přístupy. Prvním z nich je celoplošná sklizeň, která reprezentuje sklizeň všech webových stránek dostupných na doméně .cz, URL adresy webů poskytuje společnost CZ.NIC na základě smluvních podmínek. Cílem je vytvořit obraz celého českého webu v době sklizně. Tento typ sklizně probíhá jednou nebo dvakrát ročně a z kapacitních důvodů i kvůli snaze rychle zachytit obraz domény v co nejkratším čase jsou stránky sklízeny do menší hloubky.

Druhým způsobem akvizice je výběrová sklizeň. Ta obsahuje zdroje vybrané kurátory a na rozdíl od celoplošných sklizní je důraz kladen na zachycení změn zdroje v celém jeho rozsahu. Weby zařazené do tohoto typu sklizně jsou sklízeny více do hloubky a mohou se nacházet na jakékoli doméně. Musí však splňovat některé z následujících kritérií: dokument byl publikován na území České republiky, je psán v češtině, má českého autora nebo se jeho obsah týká České republiky. Preferovány jsou zdroje s dlouhodobější kulturní, vědeckou nebo historickou hodnotou.

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.[29] Specifickou podkategorií tematických sklizní jsou tzv. kontinuální sklizně. Jedná se o dlouhodobé sklízení dat v krátkých intervalech (několikrát denně) a s použitím automatizovaných postupů, kdy operátor sklizeň nespouští ručně. Tato sklizeň umožňuje sbírat data brzy po jejich zveřejnění, monitorovat průběh sklizně a vývoj tématu v reálném čase. Tímto způsobem Webarchiv sleduje v současnosti šíření a dopady viru COVID-19 na různé oblasti lidského života nebo sklízí vybraná webová periodika.

[27] Zákony pro lidi, „Zákon č. 121/2000 Sb.: Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon)."

[28] (1) Do práva autorského nezasahuje knihovna, archiv, muzeum, galerie, škola, vysoká škola a jiné nevýdělečné školské a vzdělávací zařízení, a) zhotoví-li rozmnoženinu díla, která neslouží k přímému nebo nepřímému hospodářskému nebo obchodnímu účelu, pro své archivní a konzervační potřeby, a to v počtech a formátech nezbytných pro trvalé uchování díla.

[29] Jaroslav Kvasnica, „Budoucnost českého webového archivu," in *Inforum 2015: 21. ročník konference o profesionálních informačních zdrojích* (Praha: Albertina icome Praha, 2015).

3.3 Právní rámec pro zpřístupňování dat českého Webarchivu

České právní prostředí umožňuje vytváření sbírek webového obsahu pro konzervační a archivační účely, kopie webových stránek ale mohou být zpřístupňovány pouze v prostorách budovy Národní knihovny ČR. Přestože Webarchiv sklízí pouze veřejně dostupný obsah, autorskoprávní legislativa nedovoluje zpřístupnění archivních kopií mimo prostory knihovny bez souhlasu vydavatele. Proto se kurátoři snaží alespoň část dat licenčně ošetřit, aby je Webarchiv mohl nabídnout veřejnosti volně na internetu. S vydavateli vybraných zdrojů buď uzavírají licenční smlouvu,[30] druhou možností je vystavení webu pod některou z licencí Creative Commons.[31] Volně dostupná část však představuje jen zlomek celého archivu. Jistým příslibem k širšímu zpřístupnění dat je tzv. trojnovela.[32] Vedle formulování požadavků na povinný výtisk elektronických dokumentů patří k jejím cílům i právní úprava web harvestingu a požadavek na zpřístupnění dat v několika dalších knihovnách, které jsou zákonem taxativně vyjmenované jako příjemci povinného výtisku. Z hlediska pokročilejší práce s daty je důležitá směrnice o autorském právu na jednotném digitálním trhu[33] a její budoucí implementace do české legislativy,[34] která řeší výjimky v oblasti data miningu pro paměťové instituce a bude důležitá pro formulování režimu, v jakém budou badatelé s daty pracovat. V současnosti se tedy uživatelé dostanou volně k datům zřejmě pouze v případě amerického Internet Archive, který vzniká v jiném právním prostředí. Aby se nestaly z národních webových archivů kvůli legislativním restrikcím jen darkarchivy,[35] hledají cesty, jak uživatelům nabídnout metadata, která autorským zákonem chráněna nejsou, data sestavovaná do tematických celků nebo nástroje pro pokročilejší práci s daty napříč webovým archivem pro vědecký výzkum.

[30] Webarchiv, „Nechte se Webarchivovat!," https://www.webarchiv.cz/cs/smlouva.
[31] Soubor oprávnění, která umožňují legální sdílení a využívání autorských děl.
[32] Aplikace ODok, „Návrh zákona, kterým se mění zákon č. 257/2001 Sb., o knihovnách a podmínkách provozování veřejných knihovnických a informačních služeb (knihovní zákon), ve znění pozdějších předpisů, zákon č. 37/1995 Sb., o neperiodických publikacích, ve znění pozdějších předpisů, a zákon č. 46/2000 Sb., o právech a povinnostech při vydávání periodického tisku a o změně některých dalších zákonů (tiskový zákon), ve znění pozdějších předpisů."
[33] EUR-Lex, „Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Copyright in the Digital Single Market COM/2016/0593 Final – 2016/0280 (COD)."
[34] Evropský parlament schválil směrnici o autorském právu na jednotném digitálním trhu dne 26. března 2019.
[35] Terminus technicus – archivy dostupné pouze správcům, bez přístupu veřejnosti.

4. Potřeby uživatelů webového archivu jako východisko pro vývoj nástrojů pro práci s webovými daty

Webarchiv data především dlouhodobě uchovává, chce je ale zároveň zpřístupnit uživatelům v mezích platné legislativy, vyjít vstříc jejich potřebám a zapojit se do vědeckých výzkumů. Kdo jsou jeho uživatelé? S ohledem na regionální vymezení sbírek osloví zejména ty se vztahem k České republice. Největší skupinou jsou *individuální uživatelé*, kteří mají zájem o konkrétní informace a procházejí archivovaná data prostřednictvím webového prohlížeče. *Institucionálními uživateli* jsou subjekty, které data využívají pro svou činnost v rámci státní administrativy (policie, soudy). Třetí skupinou jsou *výzkumníci a vědci,* kteří mají zájem o pokročilejší práci a většinou nemají potřebné výpočetní kapacity. Badatelská komunita projevuje o data webových archivů stále větší zájem, zejména pak o výzkumy nad rozsáhlými soubory dat, tzv. big data. Data lze zkoumat z různých perspektiv – například z hlediska vývoje jazyka, technologie, nebo designu, badatelé se zajímají o způsoby jejich vizualizace nebo textové analýzy.

Pro hledání a vývoj adekvátních nástrojů i formulování akvizičních strategií je důležité zjišťovat, kdo jsou uživatelé webových archivů, ale také jak se jejich informační potřeby mění a jak je lze naplňovat. Možnostmi užití a výzkumu v oblasti webových archivů se zabýval Jefferson Bailey, který své poznatky prezentoval roku 2016.[36] Pokusil se definovat typologii zájmů uživatelů (například zájem o extrakci informací a jejich evidenci, klasifikaci webů, sledování proměn komunikace, internetových technologií, získávání, zpracování a analýzu dat, jazyka nebo výzkum trendů na internetu), která nastínila možné oblasti výzkumu dat, k nimž patří i oblast data miningu a digital humanities. V článku *Vědecké využití dat z webových archivů*[37] byly dosavadní výzkumy včetně Baileyho poznatků shrnuty následovně:

> Data z webového prostředí jsou vnímána vědeckou společností jako poměrně nespolehlivá a pomíjivá, mnoho vědců nemá povědomí o webových archivech a netuší, jak tato data využít. Souhlasí s významem a přínosem webové archivace pro oblast výzkumu a upřednostňují tradiční přístup k datům (prohlížení

[36] Jefferson Bailey and Vinay Goel, „Program Models for Research Services," University of North Texas Libraries, UNT Digital Library.
[37] Jaroslav Kvasnica, Barbora Rudišinová a Rudolf Kreibich, „Vědecké využití dat z webových archivů," *Knihovna: knihovnická revue* 27, č. 2 (2016): 23–34.

archivních kopií). Mají zájem o fulltextové vyhledávání i o datové sety, ale často nevědí, jak je využít, jak velký vzorek si vybrat a obsahově vymezit.[38]

Z aktuálních zahraničních průzkumů a rešerší zaměřených na potřeby uživatelů webových archivů vyplývá, že se situace mění jen pomalu. Velká část akademické obce stále neví, že webové archivy existují, jaká data mají, ani jaká data by je mohla zajímat.[39] Webové archivy se napříč mezinárodní komunitou potýkají s podobnými překážkami. Vedle legislativy se jedná o nedokonalé nástroje pro sběr i zobrazení dat, nedostačující nástroje, které by umožňovaly naplnit badatelské požadavky, potenciální nespolehlivost a neúplnost archivovaných dat, nedostatečná technická dokumentace. Webové archivy zkoumají, jaká data by výzkumníky mohla zajímat. Z nejnovějších studií zaměřených na využití dat webových archivů jmenujme například *Big Data Experiments with the Archived Web: Methodological Reflections on Studying the Development of a Nation's Web*[40] zaměřenou prostřednictvím průzkumu proměn dánské národní domény v čase nebo *The Reflection of Literary Activities in Digital Space*,[41] která seznamuje s projektem Ústavu pro českou literaturu AV ČR Český literární internet,[42] na němž se Webarchiv podílel.

5. Design badatelských přístupů – aneb jak zpřístupnit webové archivy na případech Archives Unleashed a WACloud

Přístup k datům webových archivů není snadný. Způsob, jakým lze data poskytovat, úzce souvisí i s možnostmi a způsoby jejich analýzy. Ian Milligan upozorňuje, že jde o jeden z největších pramenů pro sociální dějiny současnosti a nedávné minulosti, jenž je však dosud téměř nevytěžený. Dále poukazuje na skutečnost, že se jedná o dosud nemyslitelný objem dat ke zpracování, pro vytěžení textů a zobrazení vazeb mezi stránkami / objekty,

[38] Ibid., 23–34.
[39] Maria-Dorina Costea, *Report on the Scholarly Use of Web Archives* (Aarhus: NetLab, 2018), 1–27.
[40] Niels Brügger, Janne Nielsen, and Ditte Laursen, „Big Data Experiments with the Archived Web: Methodological Reflections on Studying the Development of a Nation's Web," *First Monday* 25, no. 3 (2020).
[41] Pavla Hartmanová and Paulina Czwordon-Lis, „The Reflection of Literary Activities in Digital Space," in *11th Conference on Grey Literature and Repositories*, ed. Hana Vyčítalová (Prague: National Library of Technology, 2018), 11.
[42] Luboš Svoboda, „Webarchiv spolupracoval na projektu Český literární internet," *E-zpravodaj Národní knihovny ČR* 8, č. 4 (2021): 6.

interpretaci těchto kontextů, pro stanovení množství a podílu témat nebo identifikaci centrálních prvků sítě. Tato data jsou jak technickou výzvou, tak i velkou příležitostí moderních dějin.[43] Na obecnější úrovni teorie vědy rezonuje apel Roye Rosenzweiga z roku 2003. Dle něho bohatství digitálních pramenů vede historiky k paradigmatickému zlomu od kultury (pramenné) nouze ke kultuře bezprecedentní hojnosti literatury a pramenů, z čehož ovšem vyplývá tlak na historickou metodu.[44] Rosenzweig nabádá historiky, aby se vrátili k práci před oddělením archivnictví a historie a zasadili se o archivaci digitálních pramenů a komplexních datových struktur. Ty jsou charakteristické mírou volatility, respektive rychlejším kolísáním relevance, protože mají kratší poločas rozpadu a zastarávání než jejich tradiční tištěné a psané ekvivalenty. Jejich čtení a formátové migrace představují „zásadní technickou výzvu".[45] Tvrdí, že pro historiky může být neustále rostoucí bohatství pramenů digitální éry i digitalizované minulosti v knihovnách osvobozených od fyzických limitů svých polic zdrcující – tedy beze změny a rozvoje heuristických metod nad prameny.[46] Toto tvrzení lze dnes rozšířit obecně i na humanitně orientované vědce. Ti již dnes musí následovat datové inženýry tvořící struktury bází bez ohledu na tradiční taxonomie teoretických a informační věd. Podstatou článku není metodologická polemika, nýbrž upozornění, že nelze odmítat technologické aspekty archivace a heuristiky digitálních pramenů, neboť jinak se technologičtí optimisté a případně inženýři ve specializovaném získávání informací mohou stát budoucími vykladači historie.[47] V současnosti velké množství digitálních dokumentů a autentických pramenů schraňují soukromé společnosti, ale i národní webové archivy, které, přestože zápasí s nedostatkem prostředků, tak do jisté míry naplňují Rosenzweigovou diskusi o potřebě archivace digitálního média. V tom spočívá výzva propojování datového inženýrství a autosémantického pořádání dat s moderními, ale i tradičnějšími náhledy na zpracování informací. Hledání ve velkých datech a jejich poskytování je stále jedním z nedořešených problémů, a to zejména u národních webových archivů, institucí bytostně spjatých s prostorem internetu, který je dnes ještě daleko rozsáhlejší, než byl v roce 2003. To však nelze ze strany humanitních

[43] Ian Milligan, „Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives," *International Journal of Humanities and Arts Computing* 10, no. 1 (2016): 78–94.
[44] Roy Rosenzweig, „Scarcity or Abundance? Preserving the Past in a Digital Era," *The American Historical Review* 108, no. 3 (2003): 739.
[45] Ibid., 740–42 a následující strany.
[46] Ibid., parafráze, 756–57.
[47] Ibid., parafráze, 758.

věd označit jen za technickou překážku, na což již upozorňuje Rosenzwieg. Jde o rozpad tradičních postupů verifikace autenticity informace, její dohledatelnosti a nových způsobů čtení.[48] Je proto zásadní tento problém uspokojivě překonat, ve společné diskusi datových poskytovatelů, informačních pracovníků a celé škály badatelů.

Jedním z nejvýznamnějších aktuálních projektů vyvíjejících nástroje pro dostupnost a využití dat webových archivů patří již zmíněný Archives Unleashed. Vývojový tým archivační platformy Warcbase (2013–2017) určené pro management balíčků webových archivů na úložišti prostřednictvím technologie Hadoop, která měla sloužit k zvýšení dostupnosti dat, v roce 2017 konceptualizoval základní metodický postup vědců pro práci s daty. Jde dle nich o tzv. FAAV cyklus o čtyřech úrovních: filtruj (*Filter*), analyzuj (*Analyze*), agreguj (*Aggregate*), vizualizuj (*Vizualize*).[49] Jde o metodickou pomůcku k vyjádření potřeb badatelů pro práci s webovým archivem. V první řadě si badatel musí umět pomocí filtrů definovat pole svého zájmu na základě metadat (typ dat, čas vzniku, typ kolekce) a katalogizovaných příznaků dat (klíčová slova, téma) a následně analyzovat výsledek. Proto je nutné data extrahovat a datově/matematicky agregovat do souborů/matic (např. skrze statistické funkce, nalézání minim/maxim/průměrů/mediánů) a dále vizualizovat do přehledových tabulek a grafů. Tento postup se opakuje iterativně s různými postupně precizovanými dotazy a odpovídá Morettiho modelu tzv. vzdáleného čtení.[50] Jde o model ideotypický, skutečné postupy vědce se mohou v jednotlivých krocích lišit dle konkrétních případů užití (use case), povahy dat a postupů badatele. V posledních letech se skupina autorů projektu Archives Unleashed rozhodla upravit svůj ideotypický procesní model na tzv. FEAV, tj. analýzu vyměnila za datovou extrakci, což je přiléhavější název, protože z vyfiltrovaného materiálu si badatel nejprve extrahuje malé vzorky dat často pomocí funkcí a filtrů odpovídajícím jeho potřebám a teprve na ně pak aplikuje agregační funkce.[51] Zásadní je také zjištění, že badatelé vyžadují zejména tři typy datových derivací, které je

[48] Ibid., 741–46.

[49] Jimmy Lin et al., „Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives," *ACM Journal on Computing and Cultural Heritage* 10, no. 4 (2017): 15–17.

[50] Dle analýzy postupu vědců Lin et al., „Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives," 15–17.

[51] Nick Ruest et al., „The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives," in *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, eds. Ruhua Huang et al. (New York, NY: Association for Computing Machinery, 2020), 3–4.

možné exportovat během agregační fáze a které se vyplatí vypočítat a standardně poskytovat. Jedná se o distribuci domén a jejich frekvenci v kolekci, dále o doménový graf – pro každou z domén v kolekci a jejich vnitřní propojení, a o extrahované texty pro každou stránku s metadatovou hlavičkou s údaji o vzniku. Jejich zkušenost ukazuje, že s těmito extrakty si badatelé vystačí jak pro počáteční, tak i pro další derivované analýzy, které vzhledem k charakteru takto získaných dat je možné provádět i na notebooku, nebo dokonce v Excelu, bez složitých cloudových výpočtů nebo pokročilé znalosti datových infrastruktur webových archivů.[52] Tím zásadně mizí badatelům bariéra v průzkumu a používání dat.

Rozvíjení datových infrastruktur webových archivů s přihlédnutím k potřebám badatelů podle nás vede k minimalizaci technických nároků na uživatele a zvýšení dostupnosti dat. Analýza způsobu poskytování dat je proto zásadní, aby se z internetového obsahu mohl stát pramen pro základní výzkum. Jak již zde však bylo ukázáno v případě Warcbase/Archives Unleashed, nejde o jednoduchý nebo přímočarý export dat, ale o sadu procesů zahrnující filtraci, extrakci/ analýzu, agregaci a export nad datovými soubory, které také kopírují výše zmíněný Rosenzweigův předpoklad pro hledání a zpřístupňování digitálních dat. Design opírající se o procesní modelování výzkumných potřeb je stěžejní pro definici možností, se kterými budou uživatelé pracovat v rámci rozhraní. Jednak to umožňuje výrazně ušetřit užití náročných systémových operací výpočetního klastru nad rozsáhlými soubory dat, vede to ovšem také k vytvoření prostředí, které zjednoduší uživatelům dosažení jejich potřeb. Jistým způsobem však design takového rozhraní limituje možnosti uživatelů na předem definované cesty a operace, výměnou za konvenientní a relativně rychlý přístup k předzpracovaným datům. V projektu Archives Unleashed autoři zdůrazňují, že konceptuálně by rozhraní, jeho nástroje a služby měly být pro uživatele co možná nejjednodušší, jeho procesy téměř „neviditelné" a neměly by vyžadovat speciální znalosti pro interakci, případně nutnost mít k dispozici celou skupinu programátorů. Ve službě AUC proponují model, v němž stačí vložit do jejich cloudové technologie svá nasbíraná „surová data", AUC je zpracuje a umožní si vybrat potřebné předdefinované výstupy, i když si jsou vědomi limitací tohoto přístupu.[53]

Před otázkou, jakým způsobem designovat rozhraní a co nejméně limitovat své uživatele, stál také český Webarchiv v roce 2018 při zahájení

[52] Ibid., 4–6.
[53] Ibid., 6.

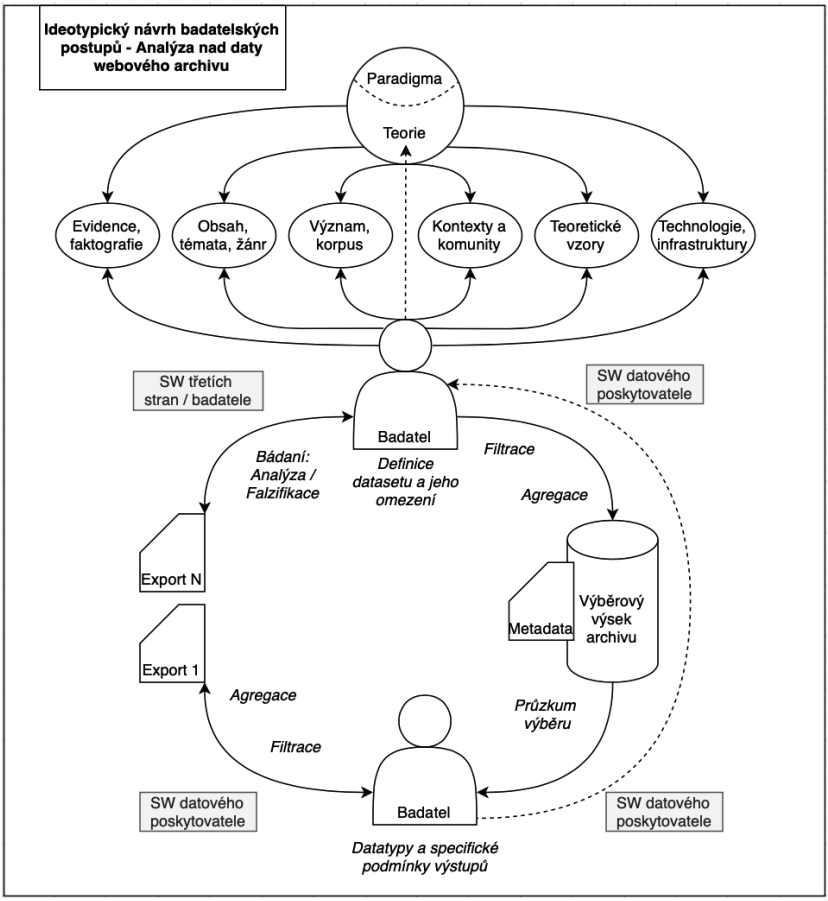
projektu Centralizovaného rozhraní pro vytěžování velkých dat z webových archivů – WACloudu. Cílem je co nejoptimálněji technologicky a uživatelsky přívětivě zpřístupnit textovou část archivovaného obsahu. Na jedné straně stojí požadavek snížení nároků na kompetence uživatele a jeho technické pochopení přípravy dat a vyčlenění datasetu, na straně druhé poskytnutí co největší variability operací a výstupů nad celým národním archivem. Při koncipování je proto nutné počítat s dostupnými prostředky softwarového i hardwarového charakteru a navrhovat robustní bezvýpadkové systémy s mohutným výpočetním výkonem založené na moderních cloudových technologiích, které je možné v budoucích letech škálovat. V případě Národní knihovny ČR byl pořízen výpočetní a úložišťový klastr.[54] I když se do velké míry iniciálně vycházelo z nastavení již zmiňovaného projektu Archives Unleashed, později Archives Unleashed Cloud,[55] v rámci projektu WACloud bylo učiněno několik rozhodnutí, která významně změnila zaměření projektu tak, aby datový archiv národní povahy mohl být použitelný v celistvosti i jeho částech.

Na vývoji rozhraní spolupracuje Národní knihovna ČR se Sociologickým ústavem AV ČR, jehož výzkumné záměry a požadavky definují část modelových postupů badatelské komunity. Další část uživatelských požadavků je formulována za účelem zobecnění pramenné základny českého Webarchivu pro další vědecké oblasti pomocí výstupů pro metodiky komputační textové analýzy (CTA), které často cílí na velká data a rozsáhlé soubory textů, procesování korpusů a přirozeného jazyka (NLP) a generování vstupů do grafové analýzy sítí, respektive sociálních sítí (NA / SNA). Důležitým partnerem projektu je Katedra kybernetiky Fakulty aplikovaných věd Západočeské univerzity v Plzni, která navrhuje řešení strojového zpracování dat, věnuje se analýze témat textového dokumentu a jejich automatické detekci na základě hlubokých neuronových sítí, rozpoznávání informací z video nebo audio souborů, tak aby bylo možné do výstupního rozhraní dodat požadovaná data a metadata. Projekt WACloud počítá s ideotypickým návrhem badatelských procesů, které nejsou platné jen pro český Webarchiv,

[54] Klastr využívá pro distribuované ukládání velkých dat oblíbené technologie Apache HDFS. Další operace probíhají již v klastru. To umožňují řešení optimalizovaná pro HDFS, jako je např. Apache Hadoop, Apache Spark a databáze Apache HBASE, které dokáží pracovat s velkým množstvím nestrukturovaných dat v řádech petabajtů. Současný klastr slouží jako komplexní *proof of case*. Do budoucna bude vhodné definovat samostatné oblasti pro ukládání, zpracování a I/O (vstupní a výstupní) intenzivní operace.

[55] Projekt Archives Unleashed se vyvíjel paralelně s projektem WACloud, na rozdíl od projektu českého Webarchivu ale AUC směřoval ke cloudové službě.

ale navrhuje tuto typizaci procesů i pro další archivy s velkým objemem dat. Výzkumník si vybere, zda chce pracovat s rozhraním v grafické podobě, nebo prostřednictvím programového rozhraní. Vytvoří si vlastní dataset, který může různě sondovat a bez nutnosti vlastních výpočetních kapacit získá výsledek komplexních analýz a agregací v podobě datového souboru pro další literární, korpusové, lingvistické, historické nebo sociologické analýzy. Předpokládáme, že metodologie pro tyto postupy se budou v nejbližších letech dále zpřesňovat. Snahou je, aby uživatelé byli co nejméně limitováni, ať už při jednoduchých operacích kvalitativního výzkumu a investigativy, nebo celých workflow pokročilých kvantitativních výzkumů. Na základě níže zobrazeného extraktivního cyklu, který umožňuje vyšší variabilitu než FEAV cyklus, může probíhat dynamická explorace a získávání přesných informací z mohutných datových souborů národních webových archivů, motivujících další výzkum, čím se otevírá cesta k řešení analytické části Rosenzweigova problému.



Obr. 1: Ideotypický návrh – Badatelská analýza nad daty webového archivu

5.1 Možné cíle analýzy a využitelnosti dat

Základní rozvrh analýzy webového archivu narýsoval Aschenbrenner a Rauber. Popisují, že mezi základní kategorie dat patří 1. sklizené stránky, 2. metadata stránek a technická metadata jejich sklízení, 3. data o užití (která ale nejsou webovými archivy získávaná standardně), 4. infrastrukturní data

o směrování a technologiích, 5. další data komplexních formátů a databází.[56] Mezi možné cíle analýzy navrhují: 1. analýzu obsahu webů a jejich technologií (kde se flexibilně prolíná užití nejen dat, ale i jejich metadat a technologií), 2. průzkum webových komunit s ohledem na různá témata a změny/progres v čase jednotlivých archivovaných snímků (umožňující například klastrování uživatelů a potažmo celé společnosti), 3. teoretický výzkum sítí (NA) a sítě v kontextu teorie grafů včetně mechanismů sdílení, sebeorganizace jejich prvků vůči entropii a 4. vzory rozvoje internetové infrastruktury.[57] Ve výše navrhovaném konceptu procesů jsme dále tato témata i s ohledem na novější Baileyho typologii začlenili jako jednotlivé prvky teoretických předpokladů metodologií do kategorií: 1. Evidence a faktografie, 2. Obsahů, témat a žánrů, 3. Významů a korpusů, 4. Kontextů a komunit, 5. Teoretických vzorů a 6. Technologií a infrastruktur, jejichž výsledky mohou mít následně zpětný vliv na výchozí teorie a paradigmata.

Pro pochopení aktuálních možností dat webových archivů s ohledem na jejich realizaci v badatelských šetřeních je např. dle Aschenbrennerovy a Rauberovy typologie nutné vést průzkum dat uložených u jednotlivých institucí. V případě českého Webarchivu statistická analýza dat ukázala, že více než polovina objemu Webarchivu je tvořena textovými informacemi, přičemž obrazová část tvoří cca 16 %.[58] Ke sklizním se evidují systematické logy soustavně od roku 2013. Výše zmíněná analýza poukázala nejen na složení archivu, co se týče mediálních typů potenciálně vhodných ke zpracování a jejich objemového i numericky absolutního zastoupení v rozsáhlém objemu dat, ale také na nízké zastoupení licencovaného obsahu, který je možné dále zpřístupňovat,[59] což komplikuje badatelskou práci. Ke dvěma základním konceptům zpracování webových stránek patří zpracování textu a zpracování obrazu. Rozvinuté zpracování obrazu řeší např. portugalský webový archiv.[60] Z hlediska postupu a významu výsledků bylo v rámci projektu WACloud vyhodnoceno jako užitečnější a uchopitelnější nejprve

[56] Andreas Aschenbrenner and Andreas Rauber, „Mining Web Collections," *Web Archiving* (1998): 155–61.
[57] Ibid., 162–72.
[58] Kvasnica et al., „Analýza českého webového archivu," 3–21.
[59] V případě počtu objektů jde o 8,17 %, v objemu sklizených domén druhé úrovně se jedná jen o 0,2 %. Kvasnica, Prokopová, Kvašová a Vozár, „Analýza českého webového archivu," 14. Viz dále Kvasnica a Kreibich, „Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR," *ProInflow* 5, č. 2 (2013): 168–77.
[60] Dle příkladu portugalského webového archivu – Arquivo.pt, který umožňuje pokročilé hledání obrázků dle formátů, velikosti, klíčových slov, ale stále zatím nemá hledání dle kategorií (viz https://sobre.arquivo.pt/en/help/help-about-image-search).

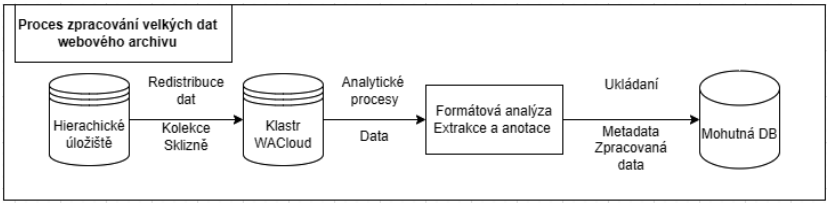
zpracování textu. Práce s analýzou a tříděním obrazových dat byla vyčleněna pro budoucí rozvoj. Koncentrace na textové materiály WACloudu vede k zpracování textů nejen ze standardního typu txt/html formátu, ale i dalších formátů, jako jsou docx, odt, pdf a aplikaci tradičních i nových metod pramenné heuristiky. Tento postup lze aplikovat i pro extrakci textů ze zvukových a audiovizuálních záznamů, což považujeme za významný přínos a rozšíření předzpracovaných dat. Pro extrakci textu z videa a ze zvukových nahrávek slouží modely za využití hlubokého strojového učení natrénované ZČU.[61] Tímto způsobem se provádí textová extrakce i nad formáty, u kterých sice existuje textuální pásmo, avšak běžnými postupy je pro většinu badatelů velice obtížně extrahovatelné.

Zásadní pro orientaci badatele je metodologie pořizování dat a jejich rozsah. Jak bylo uvedeno, Webarchiv realizuje sklizně na základě Collection policy, které podléhá obsah sklizní i jejich technické nastavení. To má významný dopad na pochopení datasetu a interpretaci výsledku badateli.[62] Obraz archivovaného obsahu je tedy nutné chápat v rámci limitací, které vedou k inherentní heterogenitě a navzdory zdánlivé hojnosti jsou nereprezentativním a neúplným výsekem z internetové sítě. S tím je nutné počítat při formulaci metod přístupů k těmto specifickým pramenům a interpretaci výsledků bádání. Limity webových archivů a jejich dopady na metodologii bádání se důkladně zaobírali v oblasti sociálních věd Matouš Pilnáček, Paulína Tabery a Martin Vávra, kteří nastínili zejména nutnost pochopení omezení jednotlivých typů sklizní limitujících dataset vždy určitým způsobem, přičemž každá má své výhody i nevýhody.[63]

[61] Zde patří poděkování zejména Pavlu Ircingovi a Janu Lehečkovi za vytvoření a implementaci postupů celkového zpracování a extrakce a též natrénování specifických modelů v nich užitých. Pro zvukovou analýzu se primárně využívá vstupní formát WAW, do kterého je nutné další populární formáty převést (viz více NAKI-NK-AUDIO, http://www.kky.zcu.cz/cs/sw/naki-nk-audio). Na tento formát je následně aplikován model hlubokých neurálních síti vytrénovaných pomocí state-of-art frameworku waw2vec 2.0 (viz více Alexei Baevski et al., „wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," preprint). Celý postup je prováděn ve škálovatelném kontejnerizovaném prostředí běžícím na on premise výpočetním klastru v Národní knihovně.

[62] Tato nastavení určují, do jaké hloubky se daný web sklízí, jak intenzivně, dlouho a jakým způsobem sklízeč ukládá kontext, zejména obsah mimo doménu.

[63] Matouš Pilnáček, Paulína Tabery a Martin Vávra, „Webové archivy a sociální vědy: příležitosti, problémy a řešení," *Naše společnost* 17, č. 1 (2019): 50–54.



Obr. 2: Proces zpracování velkých dat webového archivu

Pro poskytování obsahu webových archivů badatelům je nutné vystavět rozhraní umožňující pracovat s jeho jednotlivými částmi i celkem. Významnou vlastností rozhraní WACloud je na rozdíl od AUC holistická strategie, kdy je základním úkolem zpřístupnění dat celého národního archivu.[64] Tedy možnost pracovat potenciálně s jeho celým obsahem, s jeho ad hoc vyfiltrovanými částmi v jednom celku nebo s částmi dle výběru na základě specifických metadat. Je proto důležité, aby byl badatel informován, s jakými typy kolekcí a s kterými sklizněmi pracuje a jaké spouští extraktivní operace. Data vzniklá dle odlišných politik mohou způsobit vznik heterogenních datasetů a jejich exportů. Badatel musí být stále informován, aby mohl data vyřadit ze svého výběru tak, aby obdržel jen co nejkoherentnější dataset s ohledem na svůj výzkum. Musí mít co nejpřesnější informace o původu dat, aby se mohl kvalifikovaně rozhodnout a odhadnout dopady filtrace s ohledem na svůj badatelský záměr.

Jako příklad může posloužit zahájení operací nad kontinuální sklizní zaměřené na COVID-19, která každý den[65] do Webarchivu přidává aktuální data k tematice světové pandemie a pravděpodobně obsahuje nejucelenější soubor dat zachycující její průběh a reakce české společnosti v prostředí

[64] Zaměření AUC bylo koncipováno jako služba pro badatele, kteří si do ní mohou vložit ke zpracování vlastní data a ty zpracovat jako celek. Dle posledních informací byla 31. června 2021 služba AUC odstavena pro veřejnost (viz https://archivesunleashed.org/cloud). Dle Samanthy Fritz se projekt na základě informací z datathonů rozhodl dále integrovat ve své druhé fázi se službou Archive-It, poskytovanou Internet Archive, která provádí výběrové sklizně na požádání. Následně připraví formy agregací, které se budou pravděpodobně blížit designu systému WACloud. Samantha Fritz, „Archives Unleashed and Archive-It to Support Web Archival Research at Scale."

[65] Operativní sklízení v rámci periodicity několikrát za den ve spádové hodiny změn obsahu majoritních mediálních webů je stále v pilotní technologické fázi. Pro některé operační měsíce je dostupná jen měsíční sklizeň.

českého internetu.[66] Prostá extrakce nadpisů stránek nebo ještě lépe analýza sentimentu textu stránek spolu v kombinaci s jejich další kategorizací může vést k zajímavým výsledkům a (ne)čekaným faktickým podkladům pro analýzu chování i politik důležitých aktérů této významné etapy na úrovni velkých dějin i regionální a urbánní historie. Pokročilé filtrační mechanismy postavené za pomoci strojového učení dále umožní filtraci jenom určitých, např. reakčních, nebo tematicky zaměřených webů. Takovýto postup umožní vyfiltrovat určitá témata nebo klíčová slova a ty analyzovat na úrovni textu nebo jejich lexikálních a síťových závislostí mezi jednotlivými stránkami a realizovat tak analýzu obsahu, průzkum komunit, jejich klastrů, ale i technologického zázemí jednotlivých účastníků informačního procesu, tj. většinu analytických operací proponovaných Aschenbrennerem a Rauberem.[67] Data mohou badatelé získat použitím standardních dotazovacích polí bez dalšího programování a mohou je exportovat přímo na místě. Výsledky mohou být dále zpracovány dalšími state-of-art nástroji v rámci jejich disciplíny.[68] Proto se na rozdíl od AUC ve WACloudu neklade důraz na vizualizaci výsledků, ale na management datového přetlaku pomocí volitelných kombinací tvorby datového základu a až následných datových extrakcí, agregací a exportů výsledků.

Rozhraní WACloud tak poskytuje na rozdíl od AUC širší možnosti filtrování a zpracování dat celého archivu pomocí škálovatelného analytického zázemí, znásobujícího kapacity analýzy nad miliardami dokumentů současně. Je budováno jako ukázka nového rozhraní národního webového archivu s možností pokročilé organizace dat u kvalitativně odlišných kolekcí a jejich kombinací. V obou případech jde o zásadní usnadnění práce individuálním zájemcům o data webových archivů nad úroveň časově náročného manuálního zobrazování URL a umožnění jejich vzdáleného čtení. Rozhraní poskytuje metadata, která usnadní nacházení skrytých pramenů a umožní vymezit, vytvořit a uchovat všechny myslitelné datasety. To povede

[66] Marie Haškovcová a Zdenko Vozár, „Tematická kolekce webových zdrojů COVID-19 jako součást Webarchivu," *Bulletin SKIP* 29, č. 1 (2020).
[67] Aschenbrenner and Rauber, „Mining Web Collections," 162–72.
[68] Jde o nástroje třetích stran, nebo i vlastní nástroje zpracování. Pokročilí uživatelé mohou vybudovat v kontrolovaném prostředí vlastní zpracování nebo vizualizaci, popř. napárovat další vzdálené zdroje a služby např. pomocí tzv. *mashupu* – integrací funkcionalit různých REST API a kombinace jejich funkcionalit. Jde o primárně preferovanou cestu do budoucna. Předpokladem je, že grafická stránka ustoupí v budoucnu do pozadí a webové archivy se stanou jak službou, tak datovým zdrojem pro další komplexnější aplikace.

k zabezpečení možnosti řízené replikace jednotlivých výzkumů a vystavení teorií zkoušce zpochybněním (falzifikaci).

6. Závěr

Webové archivy jsou pro vědce stále důležitějším informačním zdrojem a zároveň otevírají celou řadu nejen nových výzkumných otázek, ale i metodologických, koncepčních, právních kompetenčních a technických výzev, které nutí humanitní a přírodní vědy hledat odpovědi společně. Navzdory počátečním problémům se v národních webových archivech za posledních více než dvacet let podařilo shromáždit významné množství dat jinak ohrožených vysokou volatilitou digitálních pramenů. Strategie a způsob jejich shromažďování nemusí být vědcům zcela srozumitelné, což může mít závažné dopady na validitu analýz a jejich syntetizujících závěrů. Proto je klíčová dokumentace na všech úrovních činnosti Webarchivu, tedy popis kurátorské strategie budování archivu, technická dokumentace procesu akvizice včetně použitých nástrojů a otisk postupů získávání a ukládání dat do jejich metadatového popisu. Badatelé mohou pak zvážit možná zkreslení v rámci jednotlivých procesů a technologií akvizice, ukládání a strukturace dat, nebo dokonce struktury konkrétních sklizní, charakteristiky extraktivních nástrojů a jejich modelů a vyvodit důsledky pro validitu a spolehlivost svých závěrů, které budou na analýzách těchto dat stavět. Iniciální míra znalosti vzniku a užití datových transformací musí být jednoduše dostupná pro všechny badatele – jak pro explorativní průzkum dat, tak i pro konstrukci vlastních datasetů včetně komplexních filtračních podmínek. S ohledem na právní a technický rámec je pro využití jejich výzkumného potenciálu potřeba hledat další možnosti. Jednou z takových cest vedle cloudové služby Archives Unleashed, určené zejména pro analýzu dat nad ucelenými datovými soubory, je WACloud – rozhraní pro flexibilní vytěžování velkých dat z českého webového archivu NK ČR. Design procesů pro extrakci dat z webových archivů, ať už AUC nebo WACloudu se opírá o potřeby poskytnout tyto unikátní prameny co nejširšímu badatelskému okruhu – od velkých výzkumných infrastruktur, přes individuální kvantitativně zaměřené badatele až po badatele se zázemím v kvalitativních metodách. Tato prostředí a jejich nástroje vznikají na základě mezioborové spolupráce ve snaze zpřístupnit badatelům dosud neznámé a nedostupné prameny pro blízké i vzdálené čtení a aplikaci různých analytických postupů.

Cílem těchto průkopnických rozhraní je umožnit badatelům lépe rozumět datům, jejich kontextům a udržitelně si vytvářet své vlastní korpusy

(datové sety). Ty umožní analýzu digitálního habitatu člověka 21. století a jeho kulturního a historického dědictví pomocí širokého spektra pokročilých kvantitativních metod a modelování digitálních humanit (CTA, NLP, SNA, statistik) a za užití specificky designovaných kategorizačních nástrojů zpřístupní bohatství webových archivů i badatelům z tradičních humanitních oborů. Tyto nástroje jsou křižovatkou propojující techniky tematického modelování, hlubokého učení a pokročilého získávání informací s tradičními postupy katalogizace. Pomocí nich ruku v ruce s postupným rozvojem svých metodologií mohou webové archivy přispět k pochopení komplexních makroskopických fenoménů současné historie – společenských fenoménů, etologie člověka ve virtuálním prostředí, obrazu o průběhu pandemie COVID-19, rekonstrukce skutečných ale i kybernetických válek, virálních fenoménů, komunitních a mikrohistorických událostí a osobních příběhů, nebo technického rozvoje lidstva a internetu samotného. Vyřešení extrakce a poskytování textuality a obrazu na agregované úrovni vede již teď ke snížení prahu dostupnosti těchto specifických pramenů pro všechny typy badatelů a do budoucna umožní další rozvoj metodik čtení digitálních dat. Předpokládáme, že povede k odkrytí zatím neznámých fenoménů ukrytých v nečekané hojnosti dnes již petabajtů virtuálních stop z minimálně posledních dvou dekád digitální historie člověka.

Bibliografie:

Aarhaus University, School of Communication and Culture. „About WARCnet: Web Archive Studies Network Researching Web Domains and Events." Navštíveno 2. listopadu 2021. https://cc.au.dk/en/warcnet/about.

Aschenbrenner, Andreas, and Andreas Rauber, „Mining Web Collections." In *Web Archiving*, edited by Julien Masanès, 155–61. Berlin: Springer, 1998.

Aplikace ODok. „Návrh zákona, kterým se mění zákon č. 257/2001 Sb., o knihovnách a podmínkách provozování veřejných knihovnických a informačních služeb (knihovní zákon), ve znění pozdějších předpisů, zákon č. 37/1995 Sb., o neperiodických publikacích, ve znění pozdějších předpisů, a zákon č. 46/2000 Sb., o právech a povinnostech při vydávání periodického tisku a o změně některých dalších zákonů (tiskový zákon), ve znění pozdějších předpisů." 2019. Navštíveno 2. listopadu 2021. https://apps.odok.cz/veklep-detail?pid=KORNBBXEMCL0.

Arquivo.pt. „Information about the Arquivo.pt Service." Navštíveno 2. listopadu 2021. https://sobre.arquivo.pt/en.

Bailey, Jefferson, and Vinay Goel. „Program Models for Research Services." University of North Texas Libraries, UNT Digital Library. Publikováno 14. dubna 2016. https://digital.library.unt.edu/ark:/67531/metadc1477166.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. „wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." Preprint, submitted June 20, 2020. https://arxiv.org/abs/2006.11477.

Brügger, Niels, and Ralpf Schroeder, eds. *The Web as History*. London: UCL Press, 2017. https://doi.org/10.2307/j.ctt1mtz55k.1.

Brügger, Niels, Janne Nielsen, and Ditte Laursen. „Big Data Experiments with the Archived Web: Methodological Reflections on Studying the Development of a Nation's Web." *First Monday* 25, no. 3 (2020). https://doi.org/10.5210/fm.v25i3.10384.

Costa, Miguel, Daniel Gomes, and Mário J. Silva. „The Evolution of Web Archiving." *International Journal on Digital Libraries* 18, no. 3 (2017): 191–205. https://doi.org/10.1007/s00799-016-0171-9.

Costea, Maria-Dorina. *Report on the Scholarly Use of Web Archives*. Aarhus: NetLab, 2018.

Cubr, Ladislav. *Autenticita a digitální informace*. Praha: Univerzita Karlova v Praze, 2017.

EUR-Lex. „Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on copyright in the Digital Single Market COM/2016/0593 Final – 2016/0280 (COD)." 2016. Navštíveno 2. listopadu 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593.

Fritz, Samantha. „Archives Unleashed and Archive-It to Support Web Archival Research at Scale." Archives Unleashed. Navštíveno 31. prosince 2021. https://news.archivesunleashed.org/archives-unleashed-and-archive-it-to-support-web-archival-research-at-scale-30e81a41f1d3.

Haškovcová, Marie a Zdenko Vozár. „Tematická kolekce webových zdrojů COVID-19 jako součást Webarchivu." *Bulletin SKIP* 29, č. 1 (2020).

Hartmanová, Pavla, and Paulina Czwordon-Lis. „The Reflection of Literary Activities in Digital Space." In *11th Conference on Grey Literature and Repositories*, edited by Hana Vyčítalová. Prague: National Library of Technology, 2018.

IIPC. „International Internet Preservation Consortium." Navštíveno 2. listopadu 2021. https://netpreserve.org/.

Internet Archive. Navštíveno 2. listopadu 2021. https://archive.org.

ISO 28500:2009. „Information and Documentation – WARC File Format." Navštíveno 2. listopadu 2021. https://www.iso.org/standard/44717.html.

Kvasnica, Jaroslav, Andrea Prokopová, Zuzana Kvašová a Zdenko Vozár. „Analýza českého webového archivu: Provenience, autenticita a technické parametry." *ProInflow* 11, č. 1 (2019): 3–21. https://doi.org/10.5817/ProIn2019-1-2 https://doi.org/10.5817/ProIn2019-1-2.

Kvasnica, Jaroslav, Barbora Rudišinová, Marie Haškovcová, Monika Holoubková a Markéta Hrdličková. „Strategie budování sbírky Webarchivu. Aktualizované znění." Národní knihovna České republiky, 2019. https://www.webarchiv.cz/static/www/download/collection-policy.pdf.

Kvasnica, Jaroslav, Barbora Rudišinová a Rudolf Kreibich. „Vědecké využití dat z webových archivů." *Knihovna: knihovnická revue* 27, č. 2 (2016): 23–34.

Kvasnica, Jaroslav. „Budoucnost českého webového archivu." In *Inforum 2015: 21. ročník konference o profesionálních informačních zdrojích*. Praha: Albertina icome Praha, 2015. https://docplayer.cz/1001491-Budoucnost-ceskeho-weboveho-archivu.html.

Kvasnica, Jaroslav a Rudolf Kreibich. „Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR." *ProInflow* 5, č. 2 (2013): 168–77.

Kvasnica, Jaroslav, Zdenko Vozár, Marie Haškovcová a Monika Kodad Holoubková. *Metodika pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu*. Praha: Národní knihovna ČR, 2020.

Lin, Jimmy, Ian Milligan, Jeremy Wiebe, and Alice Zhou. „Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives." *ACM Journal on Computing and Cultural Heritage* 10, no. 4 (2017): 1–30. http://dx.doi.org/10.1145/3097570.

Masanes, Julien. „Web Archiving Methods and Approaches: A Comparative Study." *Library Trends* 54, no. 1 (2005): 72–90. https://doi.org/10.1353/lib.2006.0005.

Milligan, Ian. „Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives." *International Journal of Humanities and Arts Computing* 10, no. 1 (2016): 78–94.

Milligan, Ian. „You Shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure." WARCnet Papers, 2020. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf.

Národní knihovna České republiky. „Webarchiv." Navštíveno 2. listopadu 2021. https://www.webarchiv.cz/cs.

Pilnáček, Matouš, Paulína Tabery a Martin Vávra. „Webové archivy a sociální vědy: příležitosti, problémy a řešení." *Naše společnost* 17, č. 1 (2019): 43–58. https://doi.org/10.13060/1214438X.2019.1.17.495.

Rosenzweig, Roy. „Scarcity or Abundance? Preserving the Past in a Digital Era." *The American Historical Review* 108, no. 3 (2003): 739.

Ruest, Nick, Jimmy Lin, Ian Milligan, and Samantha Fritz. „The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives." In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, edited by Ruhua Huang, Dan Wu, Gary Marchionini, Daqing He, Sally Jo Cunningham, and Preben Hansen, 157–66. New York, NY: Association for Computing Machinery, 2020. https://doi.org/10.1145/3383583.3398513.

Shein, Ester. „Preserving the Internet." *Communications of the ACM* 59, no. 1 (2016): 26–28. https://doi.org/10.1145/2843553.

Schafer, Valérie, and Jane Winters. „The Values of Web Archives." *International Journal of Digital Humanities* 2 (2021): 129–44. https://doi.org/10.1007/s42803-021-00037-0.

Svoboda, Luboš. „Webarchiv spolupracoval na projektu Český literární internet." *E-zpravodaj Národní knihovny ČR* 8, č. 4 (2021): 6.

Švec, Jan, Luboš Šmídl, Jan Lehečka, Pavel Ircing a Vlasta Radová. NAKI-NK-AUDIO: nástroj pro analýzu audiosouborů. Plzeň: Západočeská univerzita v Plzni, 2020.

The Archives Unleashed Project. Navštíveno 31. prosince 2021. https://archivesunleashed.org.

The Royal Library of Belgium. „Besocial." Navštíveno 2. listopadu 2021. https://www.kbr.be/en/projects/besocial.

UK Web Archive. „SHINE." Navštíveno 27. prosince 2021. https://www.webarchive.org.uk/shine.

University of London. „Big UK Domain Data for the Arts and Humanities." Navštíveno 2. listopadu 2021. https://buddah.projects.history.ac.uk.

Webarchiv. „Nechte se Webrachivovat!" Navštíveno 2. listopadu 2021. https://www.webarchiv.cz/cs/smlouva.

Zákony pro lidi. „Zákon č. 121/2000 Sb.: Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon)." Navštíveno 2. listopadu 2021. https://www.zakonyprolidi.cz/cs/2000-121.