Webové archivy a sociální vědy: příležitosti, problémy a řešení[1]

Matouš Pilnáček, Paulína Tabery, Martin Vávra, Sociologický ústav AV ČR, v. v. i.

Web Archives and Social Sciences: Opportunities, Problems and Solutions

ABSTRACT

Over the past three decades, the internet has become an integral part of contemporary societies. Online content is growing at a tremendous scale and changing dynamically. In spite of that, social sciences and social scientists have paid little attention to the kind of account of social change the World Wide Web can provide. This article provides an introduction to the subject matter of web archives, which can serve as sources of data that help us draw a picture of the dynamic change of contemporary society and communication. It seeks to discuss the different problems faced by social scientists in utilising web archive data and to propose, or at least sketch, their solutions.

In the first section of the article, we explain the purpose of web archives and their current institutional framework both in the Czech Republic and abroad. In the second section, we discuss issues of accessing web archive data. We distinguish technological access limitations, where the researcher is faced with large amounts of data and computing requirements, legal, and ethical limitations. Legal limitations break up to ones related to copyright and personal data protection. The article emphasises that there are ethical limitations in addition to technological and legal ones, which the researcher should always keep in mind and which should be respected in his/her approaching the data. As a partial solution to data access limitations, the article proposes creating and operating an analytical interface through which researchers could obtain aggregate web archive data without direct access to primary data.

KEYWORDS:

Web archives
social sciences
ethical issues
legal issues
methodological issues

DOI LINK:

http://doi.org/10.13060/1214438X.2019.1.17.495

[1] Tento článek vznikl za podpory Ministerstva kultury ČR v rámci projektu č. DG18P02OVV016 „Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů“.

Finally, the third section of the article deals with the methodological limitations of web archive data. It primarily focuses on issues of representativeness, incompleteness and heterogeneity of such data. It presents three methods of data collection for web archiving: so-called selective, thematic and full-domain harvesting. In selective harvesting, curators decide which websites to include and, from the perspective of quantitative social research, such samples are only representative of selected sources. Thematic harvesting focuses on specific topics and, as such, provides highly representative and complete collections of data as long as the researcher focuses on any of the topics covered. However, the number of topics is extremely limited. Full-domain harvesting is the most relevant method for social researchers, yet it is affected by partial non-representativeness and incompleteness. As a partial solution to the problem of limited representativeness of full-domain harvests, the authors propose implementing weighted random sampling of web archive data. Currently, however, such a solution is faced with the absence of suitable sampling frames.

The third section concludes by focusing on the issue of interpreting web archive results. An important finding is that when studies are based exclusively on such data, the researchers do not know the intentions of the actors who created the content and can only speculate about their motivations. Here, the authors see an opportunity for integrating traditional sociological research with web archive data. Furthermore, the article stresses that observed changes to online content are based not only on changes in actors' behaviours but possibly also shifts in the population of internet users, technological innovations and, last but not least, modifications of data collection methodology. It is, therefore, important for web archives to document their data collection efforts carefully and complete any analytical interfaces they provide with a precise description of the methods available to researchers.

Poslední tři desetiletí přinesla zásadní změny v komunikačních vzorcích moderních společností. Technologie umožnily nejen zrychlit komunikaci, ale rovněž ji zásadním způsobem „demokratizovat". Zavedená institucionalizovaná mediální komunikace byla postupem času čím dál víc doplňována o produkci jednotlivců, kteří pro svá vyjádření již nepotřebovali rozsáhlou infrastrukturu, ale stačil jim počítač a připojení k internetu. Internet se stal jak nástrojem oficiální komunikace, kdy instituce dávají skrze něj k dispozici úřední informace a sdělení, tak i nástrojem soukromé komunikace, a to v podobě osobních stránek, diskusí pod články, diskusních fór nebo sociálních médií. Zároveň došlo a stále dochází k obrovskému nárůstů sdělení publikovaných prostřednictvím internetu a tím ke znásobení jeho obsahu. Způsob i frekvence komunikace se tak zcela proměnily. Jak se však ptají Schroeder a Brügger [2017], co tato překotná změna vlastně říká o soudobých společnostech? Co může internet, jeho struktura a obsah říci sociálním vědcům o společenském vývoji, proměně a povaze společenské a soukromé komunikace? A vzhledem k tomu, nakolik proměnlivý web je a kolik obsahu je pozměněno a ztraceno, jak mají sociálněvědní badatelé přistoupit ke zkoumání internetového obsahu a nakolik úspěšně se mohou s těmito jevy vyrovnat?

Z teoretického hlediska je role internetu v soudobých společnostech ještě pořád nedostatečně uchopená (viz např. shrnutí v Schroeder [2018]), ať už je to dáno přílišnou specializací jednotlivých vědních oborů, jak tvrdí Schroeder [2018], nebo jenom přílišnou novostí tématu. Přesto, že je tato oblast velmi živá a prozatím neustálená, formují ji významné teorie jako například síťová teorie moci [Castells 2009], nebo teorie mediatizace [Hjarvard 2008; Couldry, Hepp 2013; Lundby 2014].

Nejistota na půdě teorie však nebrání empirickému zkoumání celé řady dílčích fenoménů a procesů. S nástupem internetu proto někteří sociální vědci začali provádět výzkum jeho obsahu a struktury [Ackland 2013; Salganik 2017]. Vznikají tak analýzy samotných webových adres [Brügger, Laursen, Nielsen 2017; Meyer et al. 2017], kvantitativní i kvalitativní analýzy obsahu [Ackland, Evans 2017; Musso, Maccaferri 2018; Schafer 2017] a také analýzy sítí odkazů mezi stránkami [Cowls, Bright 2017; Meyer et al. 2017]. Výzkumníci jsou ovšem zároveň nuceni řešit nové problémy týkající se sběru, archivace a analýzy dat. Tyto problémy lze rozdělit na ty, které se týkají samotného přístupu k datům, ať jsou to technické možnosti sběru, ukládání a analýzy takto obsáhlých dat, nebo právní a etické aspekty související se zveřejňováním tohoto druhu dat. Druhým a neméně zásadním problémovým okruhem jsou metodologické otázky, především ohledně kvality dat a jejich interpretace. Náš článek se zaměřuje právě na jmenované problematické body, které je potřeba mít na zřeteli, pokud chce výzkumník v sociálních vědách analyzovat data z internetu, především z webových stránek. V článku postupně podrobněji představíme jak jednotlivé problémy spjaté s takovou analýzou, tak nabídneme možnosti jejich řešení.

Institucionalizace webových archivů

První z problémů týkajících se výzkumu internetového prostředí je celkem zřejmý, a to je stálost a stabilita takového prostředí, ve kterém se pohybuje nepřeberné množství subjektů/autorů, kteří mohou poměrně snadným způsobem obsah měnit nebo zcela zničit. O rychle se měnícím prostředí vypovídá například studie Agata et al. [2014], kteří zkoumali převážně v japonské webové doméně dobu „přežití" webových stránek na vzorku 10 miliónů. Zjistili, že po dvanácti letech 91 % stránek na „živém webu" neexistuje (po dvou letech už nebylo dosažitelných 47 % z nich). Pokud domyslíme důsledky, znamená to, že velká část obsahu webu existujícího v minulosti (zejména z „raného" období) je definitivně ztracena. Na stejný jev poukázala i další práce. Na základě servery automaticky ukládaných záznamů o aktivitě uživatelů AlNoamany et al. [2013] zjišťovali, co uživatelé hledají v nejvýznamnějším úložišti obsahu webu, v Internet Archive[2]. Okolo 65 % stránek, které v něm uživatelé hledali, již na „živém webu" neexistovalo.

Uvedené studie jsou relativně nedávné, ovšem proměnlivá a pomíjivá podstata internetového prostředí byla jaksi intuitivně nasnadě i bez nich. Proto se už v polovině devadesátých let, v době, kdy internet ještě nedosáhl všeobecného rozšíření ve společnostech ani těch technologicky

[2] Dostupný na https://web.archive.org.

nejvyspělejších států, začaly objevovat iniciativy usilující o archivaci jeho obsahu. Již tehdy si někteří jednotlivci i instituce, často šlo o velké knihovny při univerzitách či národní knihovny, začali uvědomovat, že bez systematického uchovávání se velká část obsahů zveřejněných na internetu rychle ztrácí. Začaly tak postupně vznikat webové archivy.

Jedním z prvních repozitářů webu byl v roce 1996 Internet Archive, který dodnes představuje nejvýznamnější instituci tohoto druhu, jak ukazuje například publikace The Web as History [Brügger, Schroeder 2017], kde většina prezentovaných analýz využívá právě tento archiv. Mezi další průkopníky patřil australský archiv PANDORA[3] nebo švédský Kulturarw3[4]. Zatímco australský a švédský webový archiv, a později stejně tak mnohé další, včetně toho českého[5] nebo slovenského[6], vznikly jako součást knihoven a zaměřovaly se primárně na uchování obsahu z domén svých států, americký Internet Archive, vzniklý jako soukromá nezisková iniciativa, byl od počátku zaměřen na web v celém jeho globálním rozsahu. Od poloviny devadesátých let do současnosti vznikly desítky dalších institucí[7] uchovávající web v různé míře[8]. Institucionalizované webové archivy tak díky šíři a zejména systematičnosti ukládání dat získaly unikátní postavení při uchovávání obsahu webových stránek, kterému se archivace dat tvůrcem nebo výzkumníkem pro konkrétní projekt nebo analýzu, jak se celkem běžně děje, nemůže rovnat. S archivováním obsahu internetu pak souvisejí další otázky, v první řadě otázky přístupů k datům, ať se jedná o technická, nebo právní a etická omezení.

Přístup k datům uloženým ve webových archivech: technická omezení, právo a etika

Weber [2018: 3] uvádí, že „webové archivy představují klíčový zdroj dat pro výzkumníky", zejména pak výzkumníky v oblasti komunikačních studií, a my můžeme dodat, že i pro sociology. Když se ovšem podíváme na to, do jaké míry jsou webové archivy, a to nejen v českých sociálních vědách, využívány, musíme nutně dojít k závěru, že jde spíše o potenciál než reálné využití. Překážky, které stojí v cestě širšímu využití dat z webových archivů, jsou v zásadě dvojího rázu: jedná se jak o technické bariéry, tak především o právní a etické komplikace, které zatím zamezují většímu zájmu o archivovaný obsah webu.

Co se týče technických omezení, v současné době práce s daty uloženými ve webových archivech není lehká a vyžaduje znalosti IT postupů a technologií, kterými většina sociálních vědců nedisponuje. Jedná se o velké množství dat, která jsou náročná jak na ukládání, tak na

[3] http://pandora.nla.gov.au

[4] http://dig-hum-nord.eu/projects/kulturarw3-the-web-archive-of-the-national-library-of-sweden/

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

[6] https://www.webdepozit.sk/

[7] Zájemce o jejich seznam je možné odkázat na přehled členů International Internet Preservation Consortium (IIPC) na stránce http://netpreserve.org/about-us/members/; případně na https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

[8] Většinou právě v rozsahu definovaném doménou určitého státu, popřípadě dalšími důležitými obsahy v jiných doménách, které lze identifikovat kupříkladu pomocí jazyka.

zpracování. Výzkumníkům nemohou stačit výpočetní kapacity obvykle využívaných počítačů a často jako řešení nestačí ani využití běžného výkonného serveru. Musí tak přikročit k distribuovaným řešením, kdy výpočty běží na více výkonných počítačích paralelně v takzvaném clusteru. To jsou ovšem kapacity, které leží zcela mimo možnosti sociálních vědců a vyžadují zapojení specialistů a specializovaných pracovišť.

V současné době se věnuje poměrně velká pozornost právě těmto technickým omezením. Obvyklým řešením je přenesení technicky náročných úkolů na stranu webových archivů, které zprostředkovávají rozhraní pro přístup k datům. Díky tomu, že se problematice celkového managementu dat věnují specializované instituce, může docházet k jejich kontinuálnímu zlepšování a výzkumníci nejsou tolik zatíženi nutností zajištění náročných technických řešení. Nicméně, i když se povede odstranit tyto technické obtíže, většímu využití dat z webových archivů stále brání právní a etické otázky [Rauber, Kaiser, Wachter 2008].

Přístup do velké většiny webových archivů je pro laickou i odbornou veřejnost velmi limitován, a tento stav trvá od začátku vytváření těchto archivů do současnosti. Právní úpravy, především autorské právo a právo na soukromí se v různých zemích liší, takže zasahují přístup k datům v různé míře, ovšem i tak se zdá, že omezený přístup je převládající normou, což z něj dělá globálně sdílený úkol k vyřešení [Brügger, Schroeder 2017; Foot, Schneider, Dougherty 2003; Rauber, Kaiser, Wachter 2008]. Tento omezený přístup většinou znamená, že kompletní kolekce jsou dostupné pouze na místě, tedy přímo na vybraných počítačích v instituci, která webový archiv provozuje. Toto řešení je používané řadou archivů, například také českým[9], slovenským[10] nebo švédským[11] webarchivem. V případě českého archivu je možné většinu archivovaných materiálů na místě pouze prohlížet a nelze si z nich vytvořit kopie pro vlastní použití. Pro výzkumníka, který chce pracovat s primárními daty, je tak extrémně těžké dokumentovat svůj postup a dokládat své závěry pomocí dat z webového archivu.

Ačkoliv je to právě autorské právo, které významně určuje limity od samého zrodu webových archivů, do popředí se v poslední době dostává právo na soukromí a ochrana osobních a citlivých údajů [Dougherty 2013; Foot, Schneider, Dougherty 2003; Rauber, Kaiser, Wachter 2008]. Právní ochrana osobních údajů je v evropském kontextu zastoupená především *Obecným nařízením o ochraně osobních údajů* (General Data Protection Regulation, GDPR)[12]. Na to je potřeba pamatovat jak v souvislosti s právem subjektů údajů na vymazání informací o nich z databází, tak i v souvislosti s narušením osobních práv třetích osob. Dougherty [2013] tento posun od autorských práv k právu na soukromí nazývá odklon od obsahu k jednotlivci. Tím se dostáváme k etickému rozměru archivace dat z webů, přičemž je důležité uvědomit si, že problémy jak právní, tak zejména etické mohou nastávat ve fázi sběru dat, katalogizace i přístupu k datům [Foot, Schneider, Dougherty 2003].

9 https://www.webarchiv.cz/cs/o-webarchivu

10 https://www.webdepozit.sk/archivy-a-katalogy-dip/spristupnovanie-archivu/

11 http://dig-hum-nord.eu/projects/kulturarw3-the-web-archive-of-the-national-library-of-sweden/

12 Nařízení Evropského parlamentu a Rady (EU) č. 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (obecné nařízení o ochraně osobních údajů).

Etické otázky v oblasti webových archivů vycházejí z obecného mnohem širšího etického rámce výzkumu prováděného na internetu, ať už jde o internet jako prostor nebo jako nástroj výzkumu [více k obecnému etickému rámci viz Buchanan, Ess 2008; Salganik 2017]. Se zvětšujícím se zájmem o výzkum internetového obsahu a se zvyšujícím se povědomím o webových archivech, zvyšuje se také tlak na to, jaká etická pravidla mají být uplatňována a jak jednotlivé výzkumné projekty posuzovat. Jedná se o interdisciplinární problém a vědci z různých tradic a přístupů se tak potřebují shodnout na tom, co jejich komunita bude považovat za eticky průchozí a vhodné. Vědci z humanitních oborů se na aktéry v internetovém prostředí dívají především jako na autory, a jejich preferencí je nepřekročit právní rámec daný autorským právem. Sociální vědci tyto aktéry pojímají spíše jako subjekty, na kterých provádí svůj výzkum, a tudíž zdůrazňují hodnotu a integritu subjektů, které je potřeba chránit, a to dle pravidel, které již znají. Datoví vědci se zase kloní k ad hoc posuzování konkrétních výzkumných projektů [Buchanan, Ess 2008; Salganik 2017]. Je nasnadě, že situace je složitá nejen v rámci diskuse uvnitř vědecké komunity kvůli odlišným přístupům různých vědních disciplín, ale i v jednotlivých zemích, kde zvyklosti a právní rámec ukotvují a umožňují různé postupy při vědeckém bádání. Posuzování etiky projektu například v Spojených státech amerických a v Norsku tak může být odlišné [Buchanan, Ess 2008].

Při stanovování širšího rámce pro přemýšlení o etických výzvách v rámci výzkumu na internetu a ve webových archivech je dobré zamyslet se nad dvěma etickými přístupy, které v této oblasti existují: deontologií a konsekvencionalismem [Buchanan, Ess 2008; Salganik 2017]. Zatímco deontologické pojetí klade důraz na povinnosti a pravidla, které musí být dodrženy, aby byl výzkumný projekt etický, konsekvencionalismus se soustředí spíše na výsledky, podle kterých projekt posuzuje [ibid.]. Jak poznamenává Salganik [2017], v praxi se vyhraněně neuplatňuje ani jeden z těchto rámců a výzkumníci spíše používají mix obou přístupů. On sám pak místo přístupu založeného pouze na pravidlech (přístup typický pro sociální vědce), anebo *ad hoc* posuzování (přístup typický pro datové vědce), navrhuje etický přístup založený na principech [ibid.: 282]. Podle něj tento obecný přístup založený na respektování základních principů, jako jsou respekt k osobám, přínosy výzkumu, spravedlnost ve výzkumu a respekt k právu a veřejnému zájmu [ibid.: 294-301], dává sociálním vědcům dostatečný prostor pro posouzení rozmanitých projektů a situací, které v digitálním světě mohou nastat. K tomu bychom mohli pro českou situaci dodat, že jako limitující by se mohla ukázat absence živé diskuse o etice, stejně jako (s tím pravděpodobně spjatá) absence oborového etického orgánu v českých sociálních vědách.

Tato obecná pravidla pak nastavují rámec zcela konkrétním etickým otázkám, které si sociální vědci musí klást v souvislosti s daty uloženými ve webových archivech. Nakolik je komunikace na webu soukromá nebo veřejná, a jak na ni tedy nahlížet [Rauber, Kaiser, Wachter 2008]? Na jak dlouho by pak měla být dostupná, navždy [Foot, Schneider, Dougherty 2003; Rauber, Kaiser, Wachter 2008]? Právě identifikace komunikace jako veřejné nebo soukromé souvisí s další otázkou, kterou Rauber et al. [2008] pokládají, a to nakolik trvalé má být uchovávání věcí zveřejněných prostřednictvím internetu, a to nejen co se týče soukromých věcí. Některé věci prostě mohou být zveřejňovány se záměrem, že jsou dočasné. Dalším otazníkem je povaha

informace, která byla získána novým, opětovným vyhledáváním. Nejde pouze o opětovné zobrazení obsahu, ale o novou kvalitu obsahu, která tímto může být vytvořena. Je nutné zamyslet se, nakolik je nový kontext vypovídající a relevantní při znovuvytvoření informace, tedy nakolik je tato informace autentická a jakou novou kvalitu kumulace informací přináší [Rauber, Kaiser, Wachter 2008; Foot, Schneider, Dougherty 2003]. Je proto cenné, že uvnitř výzkumnické komunity vznikají stanoviska k etickým problémům spjatým s výzkumem obsahu internetu, jako jsou například doporučení vytvořená Association of Internet Researchers [Markham, Buchanan 2012].

V rámci právní a etické diskuze je nejzásadnější otázkou u webových archivů, nakolik budou data výzkumníkům přístupná [Rauber, Kaiser, Wachter 2008]. Opatření, která pak lze dělat, jsou jednak na straně vyhledávání dat, jednak na straně uživatelů. Rauber et al. [2008] předkládají k zamyšlení několik možností, jak se s etickými otázkami vypořádat. Lze zvažovat například oddělení oficiálního, veřejného obsahu od diskusních fór, nebo blokování jmen atd. ze stahování webového obsahu, případně posléze z vyhledávání. Lze také přemýšlet o filtrování nebo blokaci celých částí webarchivu. Rauber et al. [2008] také upozorňují, že obsah webových archivů není homogenní a různé kategorie obsahu se z hlediska etické problematičnosti liší. Jak dále uvádějí, tento problém je možné uchopit na různých úrovních: na úrovni stránek, jednotlivých typů obsahu nebo kusů odstavců a informací. Ačkoliv to vše je potřeba zohlednit při zpřístupňování obsahu webových archivů, neznamená to, že by kvůli těmto překážkám měl zůstat jejich obsah nevyužit. Naopak, jen je tak potřeba učinit právní a etickou cestou [viz Rauber, Kaiser, Wachter 2008]. Je jasné, že velké části obsahu se vznesené otázky nebudou týkat, ale u obsahu osobní povahy je potřeba etické zásady promýšlet o to pečlivěji.

Jiným opatřením je identifikovat uživatele webového archivu, a to na místě nebo registrací. Provozovatelé webových archivů by tak měli k dispozici důležité informace o tom, kdo u nich vyhledává obsah webu, i když Rauber et al. [2008] upozorňují, že i to samo o sobě může být etickým problémem. A to například kvůli rozhodování, která data kterému žadateli zpřístupnit, a tedy kvůli potenciálu pro omezování přístupu k datům pro určité kategorie uživatelů archivů. Jako řešení etických otázek týkajících se zpřístupnění dat se proto nabízí zřízení etického orgánu přímo v rámci webového archivu, který by se na tuto komplexní problematiku zaměřoval a posuzoval oprávněnost požadavků jednotlivých badatelů.

Dalším řešením některých právních, etických i technických omezení přístupů k datům může být rozhraní, které umožňuje výzkumníkům přístup k datům z webového archivu pouze na agregované úrovni. Takové řešení výrazně snižuje technickou náročnost zpracování dat, protože skrze rozhraní si může výzkumník uživatelsky přívětivě vybrat konkrétní data k analýze a získat výsledky jednoduchých analýz či datový soubor, který je podobnější datovým maticím standardně využívaným v sociálních vědách. Zároveň díky tomu, že výzkumník má přístup k datům pouze na agregované úrovni, a webový archiv tak neposkytuje původní obsah, se částečně zjednodušuje i problematika etických a právních omezení. Tento přístup poskytování agregovaných datových souborů je ovšem vhodný pouze pro kvantitativní analýzy, a ani v rámci kvantitativního výzkumu nejde o řešení zcela ideální. Přístup pouze k agregovaným

informacím značně stěžuje interpretaci získaných výsledků a badatel má omezenou kontrolu nad procesem analýzy a zpracování dat.

Jak je vidět, archivace a zpřístupňování dat zveřejněných na internetu je živá oblast výzkumu, a diskutují se jak technická řešení, tak právní konsekvence a eticky sporné momenty. Právě kvůli možné využitelnosti tohoto typu dat pro sociální vědce, je důležité pokračovat v definicích jasných a transparentních přístupových politik [viz Rauber, Kaiser, Wachter 2008]. Vzhledem k mezinárodnímu charakteru webu by pak bylo určitě přínosné dělat taková opatření na mezinárodní úrovni, nebo v souladu či součinnosti webových archivů [ibid.], i když dosáhnout toho i pouze na národní úrovni je velmi složité.

Zatím jsme se zabývali především technickými, právními a etickými problémy souvisejícími s archivací obsahu webu. Pro výzkumníky je však stejně důležité ptát se, jaká je povaha a kvalita archivovaných dat z hlediska sociálních věd, a jaká jsou jejich omezení, pokud mají vypovídat o společnosti a jejích proměnách.

Metodologické problémy dat archivovaných ve webových archivech

Data uložená ve webových archivech mají několik vzájemně souvisejících specifik, se kterými je nutné se vypořádat. Jedná se zejména o omezenou reprezentativitu, neúplnost a vysokou heterogenitu. Současný web je natolik rozsáhlý, že jeho kompletní archivace a standardní archivní postupy jsou zcela mimo reálné technické možnosti [Mason 2007]. Je proto vždy nutné vytvořit výběr, který bude archivován. Jak tento výběr odpovídá původnímu webu je samozřejmě z hlediska kvantitativní analýzy velmi podstatné. Zároveň se ne vždy daří archivovat webové stránky z technických a kapacitních důvodů celé, a tak mohou být ve webovém archivu zachovány pouze části některých webů a webových stránek.

Výzkumníci se někdy vypořádávají s otázkou reprezentativity konstatováním, že není důvod se domnívat, že archiv má v dané oblasti systematické zkreslení [Cowls, Bright 2017: 107]. Samotné konstatování by ovšem nemělo být dostačující a je vhodné reprezentativitu archivu zdůvodnit vždy pro každý výzkumný záměr. Výběr stránek do webového archivu může být totiž ovlivněn řadou faktorů, především nastavením skriptů, které data stahují, a také individuálními rozhodnutími kurátorů, kteří archiv spravují [Cowls 2017: 231].

Kolekce v archivu lze rozdělit z hlediska způsobu sběru dat do tří typů tzv. sklizní, které se v míře vlivu lidského faktoru výrazně liší [Kvasnica et al. 2017]. Zaprvé se jedná o tzv. výběrové sklizně, které se snaží zachytit kulturně hodnotné zdroje s vysokou mírou podrobnosti. U výběrových sklizní má největší vliv na obsah sbírky kurátor, který rozhoduje, jaké zdroje do sbírky zahrnout. Autoři českého Webarchivu, který provozuje Národní knihovna, deklarují, že „Cílem (…) je vytvořit reprezentativní vzorek českého kulturního dědictví (…)" [Kvasnica et al. 2017: 3], nicméně role kurátora, který nutně dělá arbitrární rozhodnutí, je zde zcela zásadní. O reprezentativitě z hlediska kvantitativního přístupu v sociálních vědách lze v tomto případě mluvit

pouze vzhledem k vybraným archivovaným zdrojům. Výběrové sklizně jsou tak zajímavé spíše pro kvalitativně orientované výzkumníky, kterým výběrová kolekce zajistí, že vybrané stránky jsou archivovány často a s velkou mírou podrobnosti, nebo pro kvantitativní výzkumníky, kteří se zajímají právě o některé z těchto zdrojů.

Druhým typem prováděných sklizní jsou sklizně tematické [ibid.], které se zaměřují na konkrétní událost (například volby) nebo téma. Zde jsou možnosti kvantitativně nahlížené reprezentativity výrazně lepší. Sklizeň může používat jak kurátorský, tak automatizovaný přístup a dosáhnout tak do velké míry pokrytí všech webových zdrojů, které se danému tématu věnovaly. Protože jde navíc jen o omezený počet webů, je možné z hlediska úplnosti zmapovat jejich větší část a nezatěžovat tím příliš datové úložiště. Tematické sklizně jsou ale z logiky věci omezené jen na několik vybraných témat. Například v rámci mezinárodní spolupráce existuje v současnosti[13] pouze pět tematických sbírek: Olympijské hry, evropská uprchlická krize, První světová válka, Mezinárodní organizace pro spolupráci a online zpravodajství[14]. Jedná se tedy o velmi kvalitní zdroj dat, pokud se výzkumník zabývá právě jedním ze zdokumentovaných témat.

Posledním typem sklizně jsou tzv. sklizně celoplošné [Kvasnica et al. 2017], které se snaží co možná nejvíce pokrýt vytyčenou archivovanou oblast (například všechny zdroje s konkrétní národní doménou nejvyšší úrovně), byť za cenu menší míry pokrytí jednotlivých webů. Tento sběr probíhá zcela automaticky a závisí tak pouze na nastavení výchozích skriptů. Z hlediska reprezentativity by tak mohlo jít o nejlepší možný způsob, jak data sbírat: pokrytí je maximální možné a nezávisí na lidském faktoru. Jak ale upozorňují Hale, Blank a Alexander [2017], k systematickým posunům dochází i v případě celoplošných sklizní, a opakovaně se ukazuje, že zkreslení je ve směru ke známějším a významnějším webovým stránkám [Ainsworth et al. 2012; Hale, Blank, Alexander 2017; Thelwall, Vaughan 2004]. Tento druh zkreslení je zapříčiněn způsobem sběru dat, když v průběhu sklizně nejsou obvykle známy adresy všech stránek, ze kterých se data mají stahovat. Skript tak postupuje skrze vybrané výchozí domény (tzv. semínka), a zároveň stahuje weby, na které je z výchozích webů odkazováno. Snadno se tak stane, že významný web, na který je odkazováno častěji, je archivován podrobněji a ve výrazně větší frekvenci než weby menší. To demonstrují Hale et al. [2017] na případové studii webu TripAdvisor, kde v současné době nejen že není jeho archivace ani zdaleka úplná, ale navíc to, co uchováváno je, nepředstavuje náhodný výběr z celé populace stránek webu TripAdvisor, ale jde spíše o archivaci stránek popisujících a hodnotících nejprominentnější, tedy nejnavštěvovanější, nebo nejvíce hodnocená místa, na jejichž recenze je nejčastěji odkazováno. Jelikož obsah některých webových stránek se může velmi často měnit [ibid.], některé historické varianty menších webů nemusí být zachyceny vůbec.

Dalším problémem celoplošných sklizní je neúplnost stažených webů. Celoplošné sklizně jsou velice obsáhlé, a je proto nutné šetřit místem na úložišti dat. Nejsou tak stahovány obvykle všechny stránky webu. U stránek s velkou mírou produkce, jako jsou například zpravodajské

[13] Současností je míněn listopad 2019.

[14] Viz http://netpreserve.org/projects/collaborative-collections/

portály, se tak snadno může stát, že je zachycena jen výrazně menší část tvorby a případná analýza tak může být velmi omezená.

Problém neúplnosti dat se ovšem týká všech typů sklizní ze dvou významných důvodů. Zaprvé ne u všeho obsahu webových stránek je možné technicky zajistit stažení. A zadruhé významná část současného webu je generována dynamicky pro každého uživatele samostatně. Takové stránky jsou staženy jen v jedné konkrétní podobě a celková výpověď o podobě webu je tím omezena. Ačkoliv tyto problémy nemají ideální řešení, je zapotřebí věnovat patřičnou pozornost nastavení stahovacího skriptu, aby bylo zajištěno stažení maximálního množství dat v co možná nejobecnější podobě.

Částečným řešením neúplnosti a nereprezentativity dat je správný výběr dat z archivu pro danou výzkumnou otázku. Vhodný výběr dat kromě určení obsahu výzkumu může pomoci zlepšit reprezentativitu a také omezit rozsah dat a snížit tak technické a výpočetní nároky.

Výběr dat z webového archivu lze rozdělit na dva základní typy označované jako „část celku" a „celek části" [Cowls 2017]. Z hlediska sociálně-vědního výzkumníka jsou tyto přístupy blízké rozdělení na kvantitativní a kvalitativní analýzu. Pokud analyzujeme část celku, tak vybíráme z celého webového archivu data o konkrétním společenském fenoménu na základě stanovené charakteristiky. Oproti tomu, pokud analyzujeme „celek části", tak si dopředu vybereme jeden či několik webů na základě expertního posouzení a ty můžeme analyzovat kvalitativním způsobem.

Co se týče výběru dat z archivu pro kvalitativní analýzu (tedy „celek části"), záleží zejména na rozhodnutí výzkumníka, jaké weby či jejich části do analýzy zařadí. Hlavním metodologickým omezením kvalitativní analýzy může být zejména neúplnost archivu. Mohou chybět například některé části stránek důležité pro interpretaci nebo v archivu nemusí být zaznamenané všechny změny, kterými daný web prošel. V tomto ohledu může být vhodné srovnávat záznamy zkoumaných webových stránek z více zdrojů, kupříkladu kombinovat národní archiv s globálním Internet Archive. Tím může dojít k výraznému zmenšení neúplnosti záznamů.

Výběr dat z archivu pro kvantitativní analýzu (tedy „části celku") se obvykle dělá dvěma způsoby. Zaprvé pomocí domén vyššího řádu, například lze vybrat několik národních domén, nebo pouze vzdělávací doménu druhého řádu „.edu" konkrétní země [Meyer et al. 2017]. Druhým způsobem je hledání stránek pomocí klíčových slov ve fulltextovém vyhledávání. Současným specifikem fulltextových vyhledávačů v akademických webových archivech je, že jsou velmi jednoduché. Nemůžeme tak očekávat kvalitu vyhledávání na úrovni, kterou poskytuje například nejznámější fulltextový vyhledávač Google. Například vyhledávač webového archivu Velké Británie Shine nedokáže řadit zdroje podle relevance, ale pouze podle času sběru či názvu položky [Cowls 2017: 235]. To může ztížit analýzu ve chvíli, kdy chceme vytipovat na základě fulltextového vyhledávání jen několik nejdůležitějších webů. Na druhou stranu, pokud je výzkumník nucen ručně projít větší množství webů, získá velmi dobrou představu o tom, jaký materiál má k dispozici a nespoléhá se na vyhledávací algoritmy, u kterých nemůže zcela zdůvodnit konečný výběr konkrétních webů [ibid.].

Nabízejí se i další možnosti vyhledávání obsahu, které nejsou na základě znalostí autorů článku dosud využívány. Například vyhledávání podle konkrétních typů webů (třeba pouze e-shopy) nebo vyhledávání v tématech přiřazených stránkám automatizovaně. Oba tyto způsoby výběru webových stránek z archivu vyžadují přístup identifikace typů stránek pomocí strojového učení a v současné době nejsou dle zkušeností autorů analytickými softwary webových archivů podporovány.

Z hlediska sociálněvědního výzkumníka je výběr z archivu v podstatě vymezením „populace" zkoumání. Kvantitativní sociologové jsou zvyklí často z dané populace udělat pravděpodobnostní výběr a zobecňovat pomocí inferenční statistiky z výběru na celou populaci a tím si výrazně zjednodušit náročnost výzkumu. Podobně je to možné udělat i u „populace" webů. Pokud provádíme obsahovou analýzu všech webů zabývající se konkrétním tématem, nemusíme nutně analyzovat všechny stránky, ale lze z nich udělat jen pravděpodobnostní výběr. Pokud jsme při procesu výběru schopni určit míru reprezentativity archivovaných stránek vůči cílové „populaci" vzhledem k nějakým proměnným (například tušíme, že máme podreprezentované menší weby), můžeme při výběru využít váhy jednotlivých jednotek tak, aby výběr byl více reprezentativní. [Kim, Wang 2018]. Kvantitativním sociologům známé postupy výběru a vážení dat tak mohou dobře sloužit ke zkvalitnění analýzy dat z webových archivů. Tento způsob zvýšení reprezentativity je velmi dobrým řešením, ale naráží na omezené možnosti dat, které mohou sloužit jako opora výběru. Informace o tom, jak je vybraná internetová „populace" stránek reprezentativní (z hlediska pro výzkum podstatných proměnných) je málokdy dostupná. Vytváření vhodných opor výběru se tak jeví jako důležitá výzva pro budoucí metodologický výzkum.

Dalším výrazným problémem dat z webových archivů je jejich vysoká heterogenita. Zaprvé se stránky liší v prezentovaném obsahu, kdy na webu nalezneme mimo jiné zpravodajské servery, e-shopy, diskusní fóra, webové rozcestníky, blogy, prezentační weby a mnoho dalších. Zadruhé se pak odlišují ve způsobu technické realizace, tedy rozložení obsahu na stránce, využití interaktivních prvků, použití videí a podobně. Může tak být extrémně obtížné transformovat data do jednotné formy tak, abychom mohli provádět kvantitativní analýzu [Hale et al. 2017: 58]. Situace se navíc komplikuje v případě longitudinální analýzy, která se u historických dat přímo nabízí. Potíž spočívá v tom, že většina stránek se v průběhu let velmi výrazně mění nejen z hlediska obsahu, ale hlavně z pohledu technické realizace.

Jako řešení vysoké heterogenity dat se nabízí redukce internetových stránek pouze na podstatné části s ohledem na analytický záměr. Většina existujících analýz se tak zabývá zejména analýzou textu [Ackland, Evans 2017; Musso, Maccaferri 2018; Schafer 2017] či hyper-textových odkazů [Cowls, Bright 2017; Meyer et al. 2017]. Extrakce odkazů z webové stránky je vzhledem k jejich jasné HTML syntaxi relativně jednoduchá. Získání textu ze stránky a odstranění textů nenesoucích význam (jako menu, patička či reklama) je technicky náročnější, ale existují již nástroje řešící tuto otázku s vysokou mírou úspěšnosti [Endrédy, Novák 2013; Pomikálek 2011].

Popsaná metodologická omezení, tedy problém reprezentativity, neúplnosti a heterogenity dat, musí mít výzkumník na paměti při celém procesu analýzy dat a interpretace výsledků. V tomto smyslu nejsou webové archivy odlišné od jakéhokoliv jiného datového zdroje v humanitních a sociálních vědách. Je proto velmi důležité, aby webové archivy vedly podrobnou dokumentaci sběru dat pro jednotlivé typy sklizní a časové období. Pokud se například zvětší rozsah celoplošné sklizně a výzkumník pozoruje nárůst objemu libovolného jevu, je nutné rozlišit, zda se jedná o reálný nárůst, nebo zda ho způsobila změna metodologie sběru. Ještě důležitější je udržovat dokumentaci metodologie pro případné aplikace poskytující přístup k agregovaným výstupům. Výzkumník musí mít dobrou představu o tom, jak proběhl výběr a zpracování dat v aplikaci, aby mohl výsledky kompetentně interpretovat.

Interpretaci výsledků stěžuje i skutečnost, že výzkumníci musí často pouze spekulovat nad důvody jednání osob, které obsah vytvářely [Salganik 2017: 87–89]. Neznáme jejich motivace ani postoje a často si nemůžeme ani být jisti tím, jaké publikum se jim podařilo zasáhnout. Například Dougherty [2017] ve své analýze islámského hnutí Taqwacore sice registruje jeho postupný úpadek na webu, ale důvody tohoto jevu analyzované pouze na základě dat z webového archivu jsou nejasné.

Oproti tradiční analýze mediálního obsahu přibývá ještě specifikum dynamického technologického vývoje. Ten nekomplikuje jen analýzu a sběr dat, jak bylo zmíněno dříve, ale také interpretaci výsledků. Například zmíněné opuštění webových stránek hnutím Taqwacore nemuselo být způsobeno zánikem komunity jako takové, ale jejím přesunem na sociální sítě se vznikem webu 2. 0. Obecně může být pozorovaná změna na webu způsobena třemi typy důvodů: zaprvé to může být tzv. populační posun, kdy se mění uživatelé webu, zadruhé může jít o posun v chování uživatelů, a za třetí může jít o změnu systému či technologickou změnu [Salganik 2017: 33]. Při interpretaci longitudinálních analýz webu je proto vždy nutné vědět, jak se proměňovali tvůrci webu a jak se měnila využitá technologie, aby bylo alespoň částečně možné odlišit pozorovanou změnu v chování. Při využití dat z webu je potřeba nezapomínat ani na „klasická" data a metody sociálních vědců – především dotazníková šetření a hloubkové rozhovory, které mohou prohloubit, či někdy vůbec umožnit interpretaci skrze porozumění aktérům.

Diskuse a závěr

Poslední tři dekády jsou ve znamení rozmachu internetu a internetové komunikace, a rovněž snahy zachytit a uchovat efemérní svět webových stránek. Můžeme tak sledovat rozvoj iniciativ zabývajících se archivací webu. Vzhledem k oddělené aktivitě archivářů a výzkumníků není prozatím plně využit potenciál webových archivů výzkumníky ze sociálních a humanitních oborů, nicméně aktivita na webu má rozhodně co říci o vývoji společnosti.

V textu jsme se věnovali problémům, které vznikají při zachovávání obsahu internetu v podobě využitelné pro vědecký výzkum. Metodologické otázky, které vyplývají z charakteru

a rozsahu dat dostupných na webu, společně s legislativou a etikou týkající se především autorských práv a ochrany osobních údajů, dávají vzniknout nejdůležitějším omezením archivace webu a jeho zpřístupňování, se kterými je potřeba se vyrovnávat. V naší studii předkládáme či alespoň nastiňujeme možná řešení uvedených problémů.

Jaké jsou tedy nejvíce aktuální problémy k vyřešení stojící před archiváři a výzkumníky? Největším problémem je zcela jistě velmi restriktivní autorské právo, které zejména v českém prostředí v podstatě znemožňuje práci s primárními daty. Tento problém je sice částečně řešitelný pomocí centralizovaného rozhraní, které dává přístup pouze k agregovaným údajům, ale to je řešení vhodné pouze pro kvantitativní výzkumníky a i tak má značná omezení. Do budoucna by proto bylo na zvážení, zda neupravit legislativu tak, aby pro základní výzkum stanovila jisté výjimky a umožnila tak badatelům pracovat s primárními daty.

Druhým významným problémem jsou nejasné etické hranice práce s daty z webových archivů. Zde se jako nejvhodnější řešení nabízí zřizování etických orgánů, a to ať už obecného rázu zabývajících se sociálněvědním výzkumem v celé jeho šíři, nebo ještě lépe orgánem přímo spjatým s konkrétním webovým archivem, který by dokázal odborně posuzovat takto komplexní problematiku v konkrétním kontextu.

V textu jsme věnovali pozornost také důležitým metodologickým otázkám kvality dat. Jako dva nejvýraznější problémy vidíme otázku neúplnosti a nereprezentativity dat. Neúplnost dat může být zásadním problémem zejména pro kvalitativní výzkumníky, zatímco klíčovou otázkou pro kvantitativní analýzu je spíše jejich reprezentativita. Vzhledem k objemu dat je pro zlepšení reprezentativity schůdným řešením dobrý výběr, který pak zároveň může vyřešit i problém s výpočetní kapacitou. Pro dobrý výběr je ovšem nutné mít přehled o důležitých charakteristikách webu a jejich změnách v průběhu času. Jako důležitý další krok umožňující lepší práci s daty z webových archivů proto vidíme budování statistik popisujících web z hlediska různých charakteristik.

V oblasti webových archivů je prozatím relativně málo studií, a v podstatě všechny lze označit za průkopnické. Dalším nutným krokem je tak začít více analyzovat data z webových archivů a při konkrétní práci s daty identifikovat další omezení a problémy, a podle toho ustavit analytické postupy, ale i pravidla nakládání s daty. Výzkumníci musí z počátku akceptovat fakt, že jejich analýzy nebudou metodologicky ideální, ale právě jen skrze konkrétní projekty je možné postupně zlepšovat kvalitu ukládání a popisování dat, a tím i kvalitu sociálněvědních výzkumů v této oblasti. Nicméně potenciál, který webové archivy pro sociální vědy nabízejí, je značný a může přinést řadu poznatků o vývoji současné společnosti.

O AUTORECH

▼

Matouš Pilnáček je odborným pracovníkem Sociologického ústavu AV ČR, v. v. i., a pracuje v oddělení Centrum pro výzkum veřejného mínění. Zároveň studuje dok-

torský program sociologie na Filozofické fakultě Univerzity Karlovy v Praze. Mezi jeho výzkumné zájmy patří problematika volebních modelů, metodologie výběrových šetření, testování nástrojů dotazování a propojování různých zdrojů dat.

▼

Paulína Tabery je odbornou pracovnicí Sociologického ústavu AV ČR, v. v. i., a vedoucí Centra pro výzkum veřejného mínění. Mezi její výzkumné zájmy patří zkoumání role interpersonální a mediální komunikace v procesu formování veřejného mínění, zkoumání názorového vůdcovství, metodologie výběrových šetření, testování nástrojů dotazování a kvality dat.

▼

Martin Vávra je postdoktorandem Sociologického ústavu AV ČR, v. v. i., a pracuje v oddělení Český sociálněvědní datový archiv. Mezi jeho výzkumné zájmy patří studování hodnot ve společnosti, management sociologických dat a studium datových zdrojů pro výzkum v sociálních vědách.

LITERATURA

Ackland, R. 2013. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. London, Thousand Oaks, New Delhi, Singapore: SAGE.

Ackland, R., A. Evans. 2017. „Using the web to examine the evolution of the abortion debate in Australia, 2005– 2015.“ Pp. 159-189 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Agata, T., Y. Miyata, E. Ishita, A. Ikeuchi, S. Ueda. 2014. „Life span of web pages: A survey of 10 million pages collected in 2001.“ Pp. 463–464 in *IEEE/ACM Joint Conference on Digital Libraries*, https://doi.org/10.1109/JCDL.2014.6970226.

Ainsworth, S. G., A. AlSum, H. SalahEldeen, M. C. Weigle, M. L. Nelson. 2012. „How Much of the Web Is Archived?“ *ArXiv E-Prints*, *1212*, arXiv:1212.6177v2.

AlNoamany, Y., A. AlSum, M. C. Weigle, M. L. Nelson. 2013. „Who and What Links to the Internet Archive.“ *International Journal on Digital Libraries 14* (3–4): 101–115, https://doi.org/10.1007/s00799-014-0111-5.

Brügger, N., D. Laursen, J. Nielsen. 2017. „Exploring the domain names of the Danish web.“ Pp. 62-82 in N. Brügger R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Brügger, N., R. Schroeder (eds.). 2017. *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Buchanan, E., C. Ess. 2008. „Internet research ethics: The field and its critical issues." Pp. 273-292 in K. E. Himma, H. T. Tavani (eds.). *The handbook of information and computer ethics*. Hoboken: John Wiley & Sons.

Castells, M. 2009. Communication Power. Oxford: Oxford University Press.

Couldry, N., A. Hepp. 2013. „Conceptualizing Mediatization: Contexts, Traditions, Arguments: Editorial". *Communication Theory* 23 (3): 191–202, https://doi.org/10.1111/comt.12019.

Cowls, J. 2017. „Cultures of the UK web." Pp. 220-237 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Cowls, J., J. Bright. 2017. „International hyperlinks in online news media." Pp. 101-116 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Dougherty, M. 2013. „Property or Privacy? Reconfiguring Ethical Concerns Around Web Archival Research Methods." *AoIR Selected Papers of Internet Research 3* (0).

Dougherty, M. 2017. „'Taqwacore is Dead. Long Live Taqwacore' or punk's not dead?: Studying the online evolution of the Islamic punk scene." Pp. 204-219 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Endrédy, I., A. Novák. 2013. „More Effective Boilerplate Removal-the GoldMiner Algorithm." *Polibits* (48): 79–83.

Foot, K., S. Schneider, M. Dougherty. 2003. „Ethics of/in Web Archiving." *Conference of the Association of Internet Researchers*. Toronto.

Hale, S. A., G. Blank, V. D. Alexander. 2017. „Live versus archive: Comparing a web archive to a population of web pages." Pp. 45-61 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Hjarvard, S. 2008. „The Mediatization of Society: A Theory of the Media as Agents of Social and Cultural Change." *Nordicom Review* 29 (2): 105– 34.

Kim, J. K., Z. Wang. 2018. „Sampling techniques for big data analysis in finite population inference." *International Statistical Review* 87 (S1): S171-S191, https://doi.org/10.1111/insr.12290.

Kvasnica, J., B. Rudišinová, M. Haškovcová, M. Holoubková, M. Hrdličková. 2017. *Strategie budování sbírky Webarchivu* [online]. Národní knihovna České republiky [cit. 30. 8. 2018]. Dostupné z: https://www.webarchiv.cz/static/www/download/collection-policy-2017.pdf.

Lundby, K. (ed.). 2014. *Mediatization of Communication*. Berlin, Boston: Walter de Gruyter.

Markham, A., E. Buchanan. 2012. *Ethical Decision-Making and Internet Research* [online]. Association of Internet Researchers [cit. 24. 7. 2019]. Dostupné z: https://aoir.org/reports/ethics2.pdf.

Mason, I. 2007. „Virtual Preservation: How Has Digital Culture Influenced Our Ideas about Permanence? Changing Practice in a National Legal Deposit Library." *Library Trends* 56 (1): 198–215, https://doi.org/10.1353/lib.2007.0055.

Meyer, E. T., T. Yasseri, S. A. Hale, J. Cowls, R. Schroeder, H. Margetts. 2017. „Analysing the UK web domain and exploring 15 years of UK universities on the web.“ Pp. 23-44 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Musso, M., M. Maccaferri. 2018. „At the origins of the political discourse of the 5-Star Movement (M5S).“ *Internet Histories* 2 (1–2): 98–120, https://doi.org/10.1080/24701475.2018.1457295.

Pomikálek, J. 2011. *jusText* [software]. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics *(ÚFAL),* Faculty of Mathematics and Physics, Charles University [cit. 24. 7. 2019]. Dostupné z: http://hdl.handle.net/11858/00-097C-0000-000D-F696-9.

Rauber, A., M. Kaiser, B. Wachter. 2008. „Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda.“ *Proceedings of the 8th International Web Archiving Workshop.* Aalborg.

Salganik, M. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Schafer, V. 2017. „From far away to a click away: The French state and public services in the 1990s.“ Pp. 117-136 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Schroeder, R. 2018. *Social Theory After the Internet: Media, Technology, and Globalization*. London: UCL Press.

Schroeder, R., N. Brügger. 2017. „Introduction: The Web as History.“ Pp. 1-22 in N. Brügger, R. Schroeder (eds.). *The Web as History*. London: UCL Press, https://doi.org/10.14324/111.9781911307563.

Thelwall, M., L. Vaughan. 2004. „A fair history of the Web? Examining country balance in the Internet Archive.“ *Library & Information Science Research* 26 (2): 162–176, https://doi.org/10.1016/j.lisr.2003. 12. 009.

Weber, M. S. 2018. „Methods and Approaches to Using Web Archives in Computational Communication Research.“ *Communication Methods and Measures* 12 (2–3): 200–215, https://doi.org/10.1080/19312458.2018.1447657.