O MRTVÝCH WEBOVÝCH ZDROJÍCH. JAK IDENTIFIKOVAT A SLEDOVAT ZANIKLÝ WEBOVÝ OBSAH?

BcA. Luboš Svoboda, DiS.; lubos.svoboda@nkp.cz; (Národní knihovna ČR)
Mgr. Marie Haškovcová; marie.haskovcova@nkp.cz; (Národní knihovna ČR)

Webový archiv Národní knihovny ČR (Webarchiv) přes dvacet let sbírá, uchovává a zpřístupňuje bohemikální webové zdroje a zachycuje tak český web. Zabývá se nejen možnostmi, jak tento proměnlivý obsah co nejlépe zachytit, ale zároveň zkoumá, jak lze na mizející obsah metodologicky nahlížet. V rámci institucionálního výzkumu navrhnul automatizované řešení pro identifikaci a popis zaniklých webů – aplikaci Extinct Websites. Tento kurátorský nástroj je jedním z kroků, jak uchopit pojem mrtvý webový zdroj pro další praxi v oblasti archivace webu a pro monitorování mizející webové krajiny. V kontextu průzkumu této oblasti v zahraničních zdrojích představuje článek možnosti, jak k problematice zanikajícího webového obsahu přistupovat.

http://doi.org/10.52036/1335793X.2023.3-4.37-41

ARCHIVACE WEBU A MIZEJÍCÍ WEBOVÝ OBSAH

Od konce 90. let začaly po celém světě vznikat webové archivy, které se snaží zachytit a uchovat nestabilní a neustále rostoucí webový obsah. Se zvyšujícím se objemem dat a v kontextu změn v posledních několika desítkách let vyvstávají nové otázky, jak s těmito unikátními daty pracovat a co všechno z nich lze o proměnách prostředí internetu i společnosti vyčíst. Ať už se jedná o obsah, sledování trendů v oblasti komunikačních platforem, technologií nebo o dynamiku mizejícího a potenciálně ohroženého obsahu, na níž jsme se, jak ukážeme dále v textu, zaměřili. Zřetelně je tento jev viditelný na mizení webového obsahu v akademické sféře, kdy se autoři potýkají se zánikem odkazů na webové zdroje, kde odkazovaný obsah buď zmizel, nebo se změnil. Například podle výzkumu odkazů v článcích New York Times vydaných mezi lety 1996 a 2019 není nejméně čtvrtina z nich dostupná a více než polovina článků obsahuje alespoň jeden nefunkční odkaz. Zatímco v textech z roku 2018 bylo nedostupných 6 % odkazů, z roku 1998 se jednalo již o více než 70 % (Bowers, Stanton a Zittrain, 2021). Mizení obsahu webů však weboví archiváři vnímají v širším kontextu. Kromě časově náročné manuální kontroly archivních kopií a jejich porovnávání s živým webem, jež lze aplikovat jen na malou část archivu, se snaží hledat postupy, jak tuto činnost automatizovat. Zaměřují se na sledování stavových kódů indikujících zánik webového zdroje i na cesty, jak identifikovat změnu obsahu. Lze však s jistotou vůbec nějaký zdroj označit za zaniklý?

Pokusíme se přiblížit naše východiska, metodologicky uchopit pojem mrtvý webový zdroj a popsat naše snahy o automatizované monitorování českých webů. Výstupem tohoto řešení (aplikace) bude databáze zaniklých webů, na základě jejíž analýzy bychom byli schopni lépe popsat tendence ve webovém prostředí a zároveň předvídat obsah, který je ohrožený. Naše aktivity se budeme snažit vsadit do kontextu podobných iniciativ v zahraničí. V některých oblastech naše úvahy a postupy směřující k automatizované identifikaci zanikajícího webového obsahu se zahraničními výzkumy konvenují, v oblasti popisu tohoto typu obsahu jsme navrhli vlastní metodologický přístup. Jak si ukážeme dále, přestože je zaniklý webový obsah pro výzkum velmi zajímavý, s ohledem na jeho proměnlivost se jedná o poměrně obtížně uchopitelný jev, který se do jisté míry přesné kategorizaci vzpírá.

JAK DEFINOVAT MRTVÝ WEBOVÝ ZDROJ

Pojem *mrtvý webový zdroj*, s nímž nakládáme, není v literatuře zatím nijak přesněji definován, ačkoliv je srozumitelné, co zhruba označuje. Když se však zaměříme na popisování webu z pohledu webového archiváře, vyjevuje se nedostatečnost tohoto pojmu. Je však třeba pokusit se ho definovat, či alespoň představit pole a možné parametry, které se s webem, který „umírá“, a jeho kategorizací pojí. V odborné mezinárodní komunitě je předmětem výzkumu jev *link rot*, česky bychom řekli možná „uhnívání odkazů“. Představme si situaci, kdy uživatel klikne v odborném textu na hyperlink

a odkazovaný zdroj se mu nezobrazí, anebo se zobrazí s jiným obsahem. Většina studií, která s termínem *link rot* zachází, tak zkoumá výběry, definované celky, vzorky, kolekce, nejčastěji z akademického či novinářského prostředí. Pojem *link rot* cílí na přerušené spojení (někdy také *link death*, *link breaking*, *broken link*, *reference rot*), popisuje spíše celky zkoumané zpětně a zajímá ho vztah URL adresy a obsahu.

Pojem *mrtvý webový zdroj* z našeho pohledu neřeší hyperlinky, aby mohl vydávat zprávu o stavu odkazů třeba na webu ČT24, ale popisuje, jak vypadá stav webové stránky, kterou takto můžeme označit z pohledu praxe webového archiváře. Míří spíše ke klasifikaci, která ji umožní uchopit v kontextu celého webu. Pokoušíme se tedy definovat stav, který nám pomůže webový zdroj jako digitální entitu dále zkoumat, nehledě na kontext jeho linkovaných částí. Je nutné stanovit si kritéria, dle kterých budeme mrtvé weby určovat, abychom je mohli dále zkoumat a sledovat co možná nejvíce automatizovaně. Pojem tedy definujeme ve vztahu k databázi, či zdroji, ve smyslu knihovnickém. Mluvíme-li o *mrtvém webu*, máme na mysli *nedostupný webový zdroj*.

Co všechno tedy musí webový zdroj splňovat, aby ho bylo možné označit za mrtvý? Představme si například stránku pivo.cz, na které je fórum, kde si fanoušci vyměňují zkušenosti s vařením piva. O několik let později doménu koupí velký pivovar a použije ji jen jako přesměrování pro svou hlavní stránku. Jedná se o mrtvý webový zdroj? Je zde komplikované mluvit o webové stránce jako o celku, je proto třeba vnímat ji jako webový zdroj s určitým obsahem, který je umístěný na konkrétní doméně. Původní zdroj pivo.cz je mrtvý, přestože doména je stále funkční a přístupná. Aktuálně se však na ní nalézá webový zdroj nový. Můžeme tak mluvit o *content driftu*, neboli změně obsahu.

Podobně sporných momentů je více, problematický je například čas – za jak dlouho je možné webový zdroj, který je nefunkční, prohlásit za mrtvý? Může „obživnout“? Jaký stav má na doméně, která je přesměrována? Je mrtvý webový zdroj ten, který nemá žádný obsah a nikdy neměl? Je web, jehož základní smysl spočívá v tom, že je dynamický (formulář, fórum) živý, přestože jeho funkcionalita nefunguje? Konkrétní příklady rozebereme v pravidlech níže. Cestu, jak tento pojem uchopit, nacházíme ve snaze vnímat weby jako zdroje obsahu – když tedy mluvíme o webu, který je mrtvý, máme na mysli především to, že obsah není nikde dostupný, a pokud stejný obsah je dostupný na jiné doméně, o mrtvý zdroj se nejedná. To je rozdíl od *link rotu*, kdy naplnění jevu spočívá v tom, že odkaz nevede tam, kam má, tedy je nedostupný. Pojem *mrtvý webový zdroj* není tedy z pohledu obsahu vázaný k doméně či jinému odkazujícímu zdroji, je důležitý pro budování databáze, jejíž průzkum bude následovat po nasbírání dat.

V angličtině existuje pojem *website* a *webpage*. V češtině jsou oba schovány pod jeden výraz *stránka*, resp. *websites* jsou *stránky* a *webpage* je *stránka*, což může být matoucí. Přestože pro naši aplikaci nemusí být tento rozdíl rozhodující, protože se v ní nacházejí oba typy zdrojů, pro definici pojmu mrtvý zdroj a další nakládání s databází je nutné mít tento rozdíl na paměti. Významnou roli hraje záměr kurátora, tedy to, jaký celek považuje za zdroj. Pokud se budeme snažit dopátrat výstižné definice, můžeme říci, že označit webový zdroj za mrtvý můžeme tehdy, když víme, že je dostupný pouze z webového archivu, a to jen do té doby, než opět bude dostupný na živém webu. Navrhli jsme několik základních pravidel pro identifikaci *mrtvého webového zdroje*, které budou blíže popsány v kapitole zaměřené na nástroj *Extinct Websites*.

O mrtvém webovém zdroji můžeme mluvit, když:
- ztratí svůj obsah,
- změní svůj obsah (*content drift*),
- zanikne, tzn. zaniká doména i obsah,
- vyprší mu hosting,
- není dostupný nikde jinde než v archivu (toto pravidlo se váže ke všem předchozím).

Jako mrtvý webový zdroj neoznačujeme:
- pokud je v rekonstrukci,
- pokud jsme ověřili přítomnost obsahu pouze jednou,
- pokud je pouze přesměrovaný na jinou doménu a obsah je stejný,
- pokud je jako zdroj vnímaný celek na úrovni domény (ve smyslu *website*, tedy soubor stránek), a nedostupná je pouze jeho část (např. článek zpravodajského serveru),
- pokud je web prázdný, nic se na něm nenachází a ani dříve nenacházelo, tzn. n*ení mrtvým webovým zdrojem*, ačkoliv má nulovou informační hodnotu,
- pokud se obsah ocitne za paywallem,
- pokud je formálně narušený, ale nadále zobrazuje obsah (například zastaralá technologie u dlouho neaktualizovaných webů),
- pokud autor na webové stránce uvádí, že její provoz skončil, ale zdroj je nadále online,
- pokud je v daném čase pozastaven pouze z důvodu nezaplacení služby, ale doména ještě neexpirovala,
- pokud proběhne redesign.

Naznačili jsme možné indikace pro i proti označení webového zdroje za mrtvý, z nichž některé jsou obtížně zjistitelné. Ze zkušenosti víme, že na jedné doméně může existovat vícero mrtvých webů, nelze se tedy řídit pouze jí. Dlouhodobé sledování mrtvých webů nám může poskytnout cenné informace o proměnách webu i společenských tendencích. Díky „úhynu" webů můžeme přesněji popisovat, kvantifikovat a zkoumat webové éry, o nichž víme, že se již dějí, či udály (blogosféra se změnila v impéria sociálních sítí), a dávat zprávy o webu do kontextu dějin nejen internetu či médií, ale i do historie obecné. Můžeme potom mluvit o životnosti webových zdrojů (*web lifespan*), o trhu, tendencích, zkoumat použité technologie, určovat to, co je ohroženo, a lépe to chránit. Nejsou to běžně dostupné informace, neboť když zanikne hostingová služba, provozovatel nemá povinnost zveřejňovat počet zaniklých url či profilů, kvantifikovat obsah, který bude nedostupný, ani poskytovat informace, co se s daty dále děje. Webarchivu se daří shromažďovat velké korpusy zdrojů zaniklých internetových služeb v prostředí českého internetu a zčásti je před vyhynutím i archivovat, přestože jde často o mravenčí práci. Důležité je mít představu o kvantitě zdrojů zaniklé služby, a mít pro účely výzkumu shromážděnou kolekci jejích zdrojů. V posledních letech se podařilo získat stovky tisíc *semínek* (URL adres) z již zaniklých služeb sweb .cz, blog.cz a webgarden.cz, které poskytovaly webové hostingy zdarma. Právě tyto zdroje patří mezi aktuálně ohrožené. V souvislosti s ochranou ohroženého obsahu bychom mohli zmínit i průzkum francouzské národní knihovny zaměřený na webové zdroje vztahující se k volbám (Amiri a Camile, 2018), z nějž vyplývá, že velká část stránek po volbách rychle zmizí – 20 % hned během následujícího roku, 56 % v průběhu následujících pěti let. Tento typ zdrojů patří i k našim prioritám.

Pokud bychom tedy dokázali sledovat úhyn webů, v čemž by nám termín *mrtvý webový zdroj* byl nápomocen, stává se z něj pojem ústřední, protože se odvíjí od smyslu webové archivace – musíme poznat, co mizí, abychom věděli, co chráníme. To, že nedostupnost na živém webu je důležitým kritériem pro určení stavu zdroje, je zřejmé, je ale také důležité zkoumat, za jakých podmínek se to děje, abychom si mohli stanovit co nejautomatizovanější postup pro určování tohoto stavu. Ačkoliv by se zdálo, že podmínka nedostupnosti webu jinde než v archivu směřuje bádání do archivů, je nutné nejprve pomocí „trackovacího" nástroje vybudovat takovou databázi, která nám to umožní. Je tedy třeba sledovat jak živý web aktuálně, tak archivní web retrográdně a komplexní jevy „mrtvosti webu" zkoumat na obou frontách, neboť jsou neoddělitelné – weby „umírají a ožívají".

TRACKOVÁNÍ ŽIVOTNÍCH FUNKCÍ WEBOVÝCH ZDROJŮ – APLIKACE EXTINCT WEBSITES

Přestože definovat mrtvý webový zdroj tak, abychom byli schopni proces automatizovat, je obtížné, pokusili jsme se určit základní pravidla, na nichž databázi stavíme. Vycházejí z průběžného zachycování stavových kódů a metadat a jejich porovnávání v čase. Na základě rozdílů mezi porovnávanými údaji je následně určen další postup. Nedílnou součástí aplikace Extinct Websites (https://github.com/WebarchivCZ/extinct-websites) je tedy trackovací část – ta v určité frekvenci ověřuje všechny webové zdroje vložené do databáze a na základě stanovených postupů určuje, zda jsou tzv. živé nebo mrtvé. Následuje fáze, v níž je potřeba akce od webového kurátora, který zhodnotí stav zdrojů, které mu aplikace předloží ke kontrole (*quality assurance*). Díky analytické části (API) pak lze analyzovat a zveřejňovat data o mrtvých webových zdrojích. Aplikace vyhodnocuje tři stavy – *Mrtvý web*, *Živý web* a *Ke kontrole*. Živý webový zdroj je ten, který je stále dostupný na živém webu, mrtvý webový zdroj jsme se pokusili definovat výše. Stav Ke kontrole označuje, že aplikace identifikovala nesrovnalosti v automatizovaném procesu, a je nutné, aby kurátor rozhodl, jaký stav webovému zdroji přiřadí.

Aplikace shromažďuje data o semínkách (URL). Můžeme rozlišit tři hlavní kategorie – *Seed report*, *Page data* a *WhoIs data*. *Seed report* zobrazuje především stavové kódy a hloubku přesměrování. *Page data* uchovává a stahuje metadata z daného webového zdroje, především H1, H2 – tedy název a nadpisy stránky – a také metadatové prvky „author" a „keywords", tzn. informace o autorovi a klíčových slovech. *WhoIs* získává údaje ve stejnojmenném registru o registraci domény, aplikace tedy vyhodnocuje změny ve vlastnictví domény.

Skrze tuto část aplikace se tedy automatizovaně dozvíme ty nejdůležitější aspekty pro vyhodnocení stavu webového zdroje, vitalitu jeho životních funkcí – zda webový zdroj nevykazuje nějaký chybový stavový kód a zda doména není přesměrována, zda se změnil obsah nebo vydavatel webového zdroje. Aplikace pak všechny sesbírané údaje uchovává a lze je zobrazit zpětně včetně časových údajů kontroly. Tomuto procesu automatizované kontroly webového zdroje lze přiřadit frekvenci kontroly, např. 1x měsíčně. Lze ji upravovat s ohledem na to, kolik zdrojů se v aplikaci nachází, či v jaké kategorii (kolekci) v databázi jsou. Nejprve

přesměrování na jinou doménu	*50*	změna domény
každé přesměrování	*5*	přírůstek za každou rozdílnou hodnotu
page data – H1	*70*	maximální skóre – procentuálně
page data – H2	*60*	maximální skóre – procentuálně
page data – autor	*40*	maximální skóre – procentuálně
page data – popisek	*20*	maximální skóre – procentuálně
page data – klíčová slova	*30*	maximální skóre – procentuálně
page data – ostatní	*5*	maximální skóre – procentuálně
whoIs	*5*	přírůstek za každou rozdílnou hodnotu

Tabulka pro stanovení indexu mrtvosti

jsme se zaměřili na zdroje, které jsou volně dostupné na základě licence, která nám umožňuje zpřístupnění jejich archivních kopií. Je jich přes 5000 (https://www.webarchiv.cz/cs/katalog-stranek). Do budoucna bychom rádi počet zdrojů rozšířili na daleko větší vzorek, který by více vypovídal o proměnách webu.

Nejsilnějším vodíkem pro určení vitality webového zdroje je detekce HTTP stavových kódů (https://cs.wikipedia.org/wiki/Stavové_kódy_HTTP) – tedy ohlas serveru na to, jak byl zpracován klientský požadavek. Aplikace rozpoznává šest typů stavových kódů: 1xx (Informational response), 2xx (Successful), 3xx (Redirection), 4xx (Client Error), 5xx (Server Error) a tzv. kódy Heritrixu, což jsou kódy vypovídající o chybách sklízení (http://crawler.archive.org/articles/user_manual/glossary.html) webu pomocí softwaru Heritrix. Workflow pracuje s dvěma skupinami stavových kódů, přičemž každá skupina podléhá jinému scénáři, detailně popsaném na GitHubu. Abychom předešli situaci, kdy se na webu děje údržba, či má výpadek, stanovili jsme rozhodovací limit porovnání třech záchytů stavových kódů po sobě. Ze získaných dat funkcí *Page Data* a *WhoIs* se automaticky tvoří tzv. *index mrtvosti*, který jsme navrhli pro zpřesnění identifikace vitality webového zdroje. Je jedním ze způsobů, jak automatizovaně určovat stav velkého objemu webových zdrojů.

Z indexu mrtvosti webového zdroje pak plyne rozhodnutí o stavu dle následujících parametrů: *Živý web* (0–50), *Ke kontrole* (více než 50), *Mrtvý web* (více než 100). Pokud nejsou k dispozici žádná metadata, je zdroj uložen jako živý, ale je zaznamenána informace o nedostupnosti metadat. Při určování stavů webových zdrojů jsou cenná i data z archivovaných zdrojů, která mohou posloužit při identifikaci změn na doméně. Například britský webový archiv použil v roce 2015 systém fingerprintu (Jackson, 2015), který přiřadil ke každé archivované URL hash kód, do kterého se promítá proměna stránky na základě textových posunů v těle stránky. Díky tomu mohli britští weboví archiváři určit podobný systém našemu *indexu mrtvosti* pouze na základě parsování textu z archivních verzí webů v čase, a tvořit index změny, přiřadit mu definované procentuální hodnoty, a následně určit jeho pravděpodobný stav na škále změn. Pro extrakci dat o mrtvých webových zdrojích z archivních kopií vyvíjíme nástroj na analýzu heritrixových logů *Logparser.* Aplikaci je třeba podrobit dlouhodobějšímu testování a vyhodnocení výsledků.

Vzhledem k tomu, že ne všechna kritéria mohou být automaticky detekovatelná, důležitou roli v procesu určování stavů zdrojů má kurátor. Pokud aplikace označí zdroj příznakem *Ke kontrole*, kurátor musí vyhodnotit, zda se jedná například o změnu obsahu či ne. K tomu mu slouží jak porovnání archivní verze webových zdrojů s živým webem, tak historie funkcí aplikace *Page Data* a *WhoIs*. Velmi obtížné je také určit, zda webový zdroj zanikl, či se nachází na jiné doméně. V tomto případě zatím nemá kurátor příliš možností, jak stav ověřit, pokud se informace nenacházejí přímo na zkoumané stránce. Pokud nic nenasvědčuje tomu, že se web může na jiné doméně nacházet, webový zdroj je označen za mrtvý.

ZPRÁVA O MRTVÝCH WEBOVÝCH ZDROJÍCH

Aplikace Extinct Websites umožňuje pomocí API agregovat údaje o mrtvých webových zdrojích po každé kontrole dat. Výstupem bude databáze zobrazující statistiky, z nichž si bude možné udělat představu o trendech zanikajících zdrojů. Hlavní osou bude čas a počet

zaniklých webů. Do budoucna by bylo prospěšné si stanovit topologii, na níž by bylo vidět, jaké typy webových zdrojů zanikají (např. blog, fórum, sociální síť atd.) a popsat tak éry webové sféry komplexněji (např. web 1.0/2.0, poměr statického a dynamického webu atd.). Další perspektivou by mohla být analýza obsahu z hlediska typu média (text, audio, video) nebo doménové vymezení – tzn. kolik bohemikálních webů na doméně .cz/.net/.eu/.com ad. zaniklo. Formátovou analýzu jsme již v minulosti provedli na vzorku archivních verzí stránek (Kvasnica a Kreibich, 2013). Představu o struktuře archivních dat Webarchivu nabídla i analýza indexu, nebo-li seznamu veškerého archivovaného obsahu, provedená v rámci výzkumného projektu *Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů* (Kvasnica, Prokopová, Vozár a Kvašová, 2019). Zahrnovala mimo jiné analýzu zastoupení domén nebo identifikátorů souborových formátů, tzv. MIME typů. Dosavadní průzkumy však nezohledňovaly, zda se jedná o zaniklý webový obsah. Databáze může být výchozím bodem pro další výzkumy sledující například zánik celé platformy (např. blog.cz), změnu obsahu, záměrné znepřístupnění celku či části zdroje anebo ukončení vydávání zdroje jeho vydavatelem. Vodítkem k tomu může být statistika stavových kódů. Smyslem aplikace je nastavit agendu dlouhodobého pravidelného sledování zanikajících webových zdrojů tak, abychom pomocí uložených metadat o tak dynamickém médiu, jako je internet, mohli popisovat webovou krajinu.

Recenzovaný článek vznikl na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace Národní knihovna České republiky poskytované Ministerstvem kultury ČR.

POUŽITÁ LITERATURA:

AMIRI, Maral a Géraldine CAMILE, 2018. *Quelle durée de vie ont les sites sélectionnés dans les collectes du web électoral? Web Corpora: Explorer les archives de l'internet à la BnF* [online]. 22 June 2018 [cit. 2023-09-23]. Dostupné z: https://webcorpora.hypotheses.org/540, https://web.archive.org/web/20211203124742/https://webcorpora.hypotheses.org/540

BOWERS, John, Clare STANTON a Jonathan ZITTRAIN, 2021. *What the ephemerality of the Web means for your hyperlinks. Columbia Journalism Review* [online]. 21 May 2021 [cit. 2023-09-28]. Dostupné z: https://www.cjr.org/analysis/linkrot-content-drift-new-york-times.php, https://web.archive.org/web/20210521143132/https://www.cjr.org/analysis/linkrot-content-drift-new-york-times.php

JACKSON, Andy, 2015. *Ten years of the UK web archive: what have we saved? UK Web Archive blog* [online]. 18 August 2015 [cit. 2023-09-28]. Dostupné z: https://blogs.bl.uk/webarchive/2015/09/ten-years-of-the-uk-web-archive-what-have-we-saved.htm, https://web.archive.org/web/20170412004230/http://blogs.bl.uk/webarchive/2015/09/ten-years-of-the-uk-web-archive-what-have-we-saved.html

JONES, Shawn M., Martin KLEIN a Herbert Van de SOMPEL, 2021. *Robustifying Links To Combat Reference Rot. Code4Lib Journal* [online]. 10 February 2021. Code4Lib, 50 [cit. 2023-09-28]. ISSN 1940-5758 (online). Dostupné z: https://journal.code4lib.org/articles/15509, https://web.archive.org/web/20230601150447/https://journal.code4lib.org/articles/15509

KVASNICA, Jaroslav a Rudolf KREIBICH, 2013. Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. *ProInflow: Časopis pro informační vědy* [online]. 5(2), 168-177 [cit. 2023-09-28]. ISSN 1804-2406 (online). Dostupné z: https://doi.org/10.5817/ProIn2013-2-14, https://web.archive.org/web/20221221110754/https://journals.phil.muni.cz/proinflow/article/view/2013-2-14

KVASNICA, Jaroslav, Andrea PROKOPOVÁ, Zdenko VOZÁR a Zuzana KVAŠOVÁ, 2019. Analýza českého webového archivu: Provenience, autenticita a technické parametry. *ProInflow: Časopis pro informační vědy* [online]. 11(1), 3-21 [cit. 2023-09-28]. ISSN 1804-2406 (online). Dostupné z: https://doi.org/10.5817/ProIn2019-1-2, https://web.archive.org/web/20220210134810/https://journals.phil.muni.cz/proinflow/article/view/2019-1-2

LOBBÉ, Quentin, 2018. Where the dead blogs are. A Disaggregated Exploration of Web Archives to Reveal Extinct Online Collectives. *ICADL 2018 - 20th International Conference on Asia-Pacific Digital Libraries, Hamilton* [online]. November 2018 [cit. 2023-09-28]. Dostupné z: https://hal.archives-ouvertes.fr/hal-01895955/document, https://web.archive.org/web/20231011212302/https://hal.science/hal-01895955/document

MEEHAN, Olivia, 2022. *Diving into Digital Ephemera: Identifying Defunct URLs in the Web Archives. Library of Congress Blogs* [online]. 3 August 2022 [cit. 2023-09-28]. Dostupné z: https://blogs.loc.gov/thesignal/2022/08/diving-into-digital-ephemera-identifying-defunct-urls-in-the-web-archives/, https://web.archive.org/web/20230602122348/https://blogs.loc.gov/thesignal/2022/08/diving-into-digital-ephemera-identifying-defunct-urls-in-the-web-archives/

OOGHE-TABANOU, Benjamin, Mathieu JACOMY, Paul GIRARD a Guillaume PLIQUE, 2018. *Hyperlink is not dead! International conference on Web Studies, Paris* [online]. New York: ACM. 3-5 October 2018 [cit. 2023-09-28]. Dostupné z: https://doi.org/10.1145/3240431.3240434

VOZÁR, Zdenko, Marie HAŠKOVCOVÁ a Andrea PROKOPOVÁ, 2022. Internet jako pramen výzkumu: přístup k archivovaným webovým zdrojům a možnosti jejich zpracování. *Teorie vědy* [online]. 44(1), 59-87 [cit. 2023-09-28]. ISSN 1804-6347 (online). Dostupné z: https://doi.org/10.46938/tv.2022.552, https://web.archive.org/web/20221207134700/https://teorievedy.flu.cas.cz/index.php/tv/article/view/552

■ Článok bol recenzovaný