Haškovcová, Marie; Svoboda, Luboš; Hrdličková, Markéta

Používáte Webarchiv? : průzkum potřeb uživatelů českého webového archivu

ProInflow. 2022, vol. 14, iss. 1-2, pp. [4]-30

ISSN 1804-2406 (online)

Stable URL (DOI): https://doi.org/10.5817/ProIn2022-2-2
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.77643

ProInflow: časopis pro informační vědy
Vol. 14, No. 1–2, 2022, str. 4–30
https://doi.org/10.5817/ProIn2022-2-2

https://journals.phil.muni.cz/proinflow/

POUŽÍVÁTE WEBARCHIV? PRŮZKUM POTŘEB UŽIVATELŮ ČESKÉHO WEBOVÉHO ARCHIVU

DO YOU USE WEBARCHIV? SURVEY OF THE USERS NEEDS OF THE CZECH WEB ARCHIVE

Marie Haškovcová, Luboš Svoboda, Markéta Hrdličková

Národní knihovna ČR, Oddělení archivace webu

Abstrakt

Účel – Cílem dotazníkového šetření bylo zmapovat povědomí akademické sféry o existenci a využití webových archivů, zjistit, zda badatelé s daty webových archivů pracují a jak by je případně chtěli ve svých výzkumech používat. Závěry výzkumné studie budou sloužit jako podklad pro koncipování dalších výzkumných činností Národní knihovny ČR.

Design/metodologie/přístup – Pro zjištění potřeb uživatelů webových archivů byla zvolena metoda kvantitativního dotazníkového průzkumu. Vzniku dotazníku předcházela rešerše zabývající se průzkumy uživatelských potřeb v zahraničí a vymezení cílové skupiny. Průzkumu se účastnili vědci, kteří mají zkušenosti s webovými archivy, i ti, kteří se s nimi dosud nesetkali. Otázky byly strukturovány do několika oblastí zaměřených na to, kdo jsou respondenti, jaké mají povědomí o webových archivech, dále jsme se ptali na způsob a záměr používání webového archivu a na datovou gramotnost. S žádostí o distribuci dotazníku co nejširšímu okruhu respondentů z oblasti akademické sféry byly cíleně osloveny především univerzity a vysoké školy, veřejné výzkumné instituce, krajské a vědecké knihovny a členské knihovny Asociace vysokoškolských knihoven. Zpracováno a vyhodnoceno bylo všech 146 odpovědí.

Výsledky – Studie představuje výsledky v českém prostředí jedinečného průzkumu. Shrnuje obecné informace o respondentech, o jejich povědomí o webových archivech a jejich využití, o zkušenostech s prací s archivními daty i o konkrétních požadavcích na český Webarchiv. Poznatky zasazuje do kontextu zahraničních výzkumů potřeb uživatelů webových archivů a naznačuje další možnosti výzkumu těchto archivních dat.

Originalita/hodnota – Průzkum uživatelských potřeb formou rozsáhlého dotazníkového šetření, které umožnilo zmapovat používání webových archivů, se v českém prostředí uskutečnil poprvé. Přestože se webový archiv Národní knihovny ČR průběžně zabývá tím, jak může v rámci legislativy svá data a služby poskytovat uživatelům, výstupy z dotazníkového šetření poskytly unikátní poklad pro další vývoj služeb a funkcionalit.

Klíčová slova: Webarchiv, dotazník, uživatelské potřeby, archiv, badatelé, digital humanities, digitální dědictví

Abstract

Purpose – The aim of the questionnaire survey was to map the awareness of the academic sphere about the existence and use of web archives and find out how researchers would like to use archival data in their research. The conclusions of the research study will serve as a basis for designing further research activities of the National Library of the Czech Republic.

Design/Methodology/Approach – In order to find out the needs of users of web archives the method of quantitative questionnaire was chosen. The process of making the questionnaire was preceded by research that focuses on survey of user needs abroad and definition of the target group. The survey included scientists experienced with web archives as well as those who have not participated in a similar project. The questions were structured into several areas focused on who the respondents are and what is their knowledge about web archives. We also asked about the way and intention of using the web archive and about data literacy. With the request to distribute the questionnaire to the widest possible range of respondents from the academic sphere, universities and colleges, public research institutions, regional and scientific libraries and members libraries of the Association of university libraries were targeted. All of the 146 responses were processed and evaluated.

Results – The study summarizes the results of a unique survey in the Czech environment, places them in the context of foreign research on the needs of web archive users and suggests further possibilities for research on these archival data. It summarizes general information about the respondents, their awareness of web archives and its use as well as about experiences working with archive data and about requirements for the Czech Webarchive.

Originality/Value – The survey of user needs in the form of an extensive questionnaire survey happened for the first time in the Czech environment and thus made it possible to map the use of web archives. Although the web archive of the National Library of the Czech Republic continuously deals with how it can provide its data and services to users within the framework of legislation, the results of the questionnaire survey provided a unique treasure for the further development of services and functionalities.

Keywords: Webarchiv, survey, user needs, archive, researchers, digital humanities, digital heritage

ÚVOD

Webový archiv Národní knihovny ČR více než 20 let uchovává českou webovou krajinu. Kromě sběru a uchovávání webových dat se snaží svůj unikátní obsah zpřístupňovat veřejnosti. Aby mohl uživatelům vyjít co nejlépe vstříc, připravil dotazníkové šetření, jehož cílem je zjistit, jaké je povědomí akademické sféry o existenci a využití webových archivů nebo jakým způsobem by badatelé chtěli archivní data používat. Závěry šetření budou sloužit jako podklad pro koncipování dalších výzkumných činností Národní knihovny ČR. Studie shrnuje výsledky v českém prostředí jedinečného průzkumu, zasazuje je do kontextu zahraničních výzkumů potřeb uživatelů webových archivů a naznačuje další možnosti výzkumu těchto archivních dat.

Existence webových archivů je klíčová pro uchování části národního kulturního dědictví, které vzniká nebo se vyskytuje v prostředí internetu. Ať už se jedná o informace vládních webů, osobních stránek jednotlivců nebo o jevy s online prostorem neoddělitelně spjaté, jako je prostředí sociálních médií, digitálního obchodu nebo streamovacích služeb. Archivovaná data jsou cenným zdrojem informací pro vědce různého zaměření, historiky, sociology, výzkumníky v oblasti digital humanities a další, mají ale také potenciál sloužit do budoucna například jako unikátní datasety pro strojové učení. Paměťové instituce, které se archivaci webu věnují, se zprvu soustředily zejména na vývoj technických aplikací, standardů pro uchování dat a formulování akvizičních strategií, aby efemérní webový obsah dokázaly co nejlépe uchovat. Postupně svou pozornost více obracely ke svým uživatelům a začaly se zamýšlet nad tím, jak s archivními daty pracovat. Přestože se snaží obsah zachytit, uchovat a archivní data zpřístupňovat veřejnosti již řadu let (historie nejstarších sahá do konce 90. let), povědomí o jejich činnosti stále není příliš velké. Proto hledají možnosti, jak svá data co nejlépe využít. Významnou platformou pro sdílení zkušeností v různých oblastech archivace webu, včetně právních a etických oblastí, je International Internet Preservation Consortium. Tvoří ji mezinárodní komunita více než 50 institucí zaměřených na archivaci webu, usiluje o společné formulování tzv. best practices, vytváření nástrojů a standardů. Od roku 2007 je jejím členem i český webový archiv.

ARCHIVACE ČESKÉHO WEBU

Archivaci obsahu českého internetu zajišťuje od roku 2000 webový archiv Národní knihovny ČR – Webarchiv. Vznikl s cílem shromažďovat, ochraňovat, zpřístupňovat a dlouhodobě uchovávat webový obsah pro budoucí generace. Zaměřuje se na dokumenty s bohemikálním obsahem. Vymezení akviziční strategie je zásadní pro další výzkum – pro formulování badatelských záměrů i interpretaci jejich výsledků. Kritéria výběru zdrojů i právní rámec fungování archivu jsou popsány v dokumentu Collection Policy (Kvasnica, Rudišinová, Haškovcová, Holoubková & Hrdličková, 2019). Knihovní licence Autorského zákona (č. 121/2000 Sb.) umožňuje knihovně vytvářet rozmnoženiny díla, tzn. archivní kopie webů, pro své archivní a konzervační účely. Zatímco celý webový archiv je možné si prohlédnout v budově Národní knihovny na speciálně určených terminálech, mimo budovu smí zpřístupnit pouze licenčně ošetřené zdroje, což představuje méně než 0,4 % obsahu celého archivu. Strategie sběru dat kombinuje tři přístupy. Sklizně celoplošné se zaměřují na data z české domény .cz. Výběrové se soustředí na dlouhodobé sklízení hodnotných webů, přičemž výběr podléhá kurátorskému rozhodnutí na základě definovaných kritérií a zdroje jsou licenčně ošetřeny (buď uzavřením smlouvy s vydavateli nebo vystavením webů pod licencí Creative Commons). Tematické sledují určité události nebo témata (volby, válka na Ukrajině, Covid-19, olympijské hry). Data jsou uchovávána ve standardizovaném kontejnerovém formátu WARC a jsou opatřena technickými a administrativními metadaty, která se vztahují k procesu sklízení, např. datum nebo typ sklizně. Detailně jsou popsány v Metodice pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu (Kvasnica, Vozár, Haškovcová & Kodad Holoubková, 2020). Licenčně ošetřené zdroje mají kromě toho i metadata bibliografická – mají katalogizační záznamy v systému Aleph, které jsou součástí České národní bibliografie.

Český webový archiv je otevřený spolupráci s badateli a institucemi, například s Ústavem pro českou literaturu Akademie věd ČR spolupracuje na archivaci českých literárních zdrojů v rámci projektu Český literární internet (Svoboda, 2021). Jednou z cest, jak data otevřít k výzkumnému využití, je vývoj nástrojů umožňujících práci s daty pokročilejšími způsoby než jen zobrazením archivních kopií prostřednictvím zpřístupňovací aplikace. Na badatelskou komunitu tímto způsobem cílí projekt, jehož výstupem bude výzkumné rozhraní pro vytěžování velkých dat z webových archivů s využitím postupů strojového zpracování, které bude badatelům poskytovat datasety dle jejich specifických výzkumných požadavků (Vozár, Haškovcová & Prokopová, 2022). Na jeho vývoji se podílí Národní knihovna ČR, Sociologický ústav AV ČR a Katedra kybernetiky Fakulty aplikovaných věd Západočeské univerzity v Plzni.

KDO JSOU UŽIVATELÉ WEBOVÝCH ARCHIVŮ?

Podobně jako zahraniční archivy se český webový archiv zabývá nejen tím, jak data sbírat, uchovat a zpřístupnit, ale i způsoby, jakými nashromážděná data interpretovat, jak je dále v rámci platné legislativy využívat a poskytovat badatelům k dalším výzkumům. Své uživatele dělí do několika skupin. Kromě individuálních uživatelů jsou to institucionální uživatelé, kteří data používají pro svou činnost (policie, soudy) a výzkumníci a vědci, kteří mají zájem o pokročilejší práci s datovými sety a o výzkumy nad velkými objemy dat (big data).

Typologií zájmu uživatelů webových archivů se zabýval Jefferson Bailey, své poznatky prezentoval na konferenci IIPC v roce 2016 a naznačil oblasti dalšího výzkumu (Bailey & Goel, 2016). Zájmy uživatelů rozčlenil do šesti skupin: dokumentární oblast (ověřování informací, klasifikace nebo evidence stránek), dále sociální a politické výzkumy, oblast web science (internetové technologie a protokoly), digital humanities (oblast humanitních oborů, které pracují s digitálními daty), computer science (vyhledávání informací, zpracování dat, indexace, infrastruktura, nástroje) a datová analytika (vytěžování a trénování dat, zpracování jazyka, analýza trendů na internetu). Potřeby uživatelů webových archivů byly i s odkazem na Baileyho kategorizaci reflektovány v soudobé české odborné literatuře tak, že mnoho vědců o webových archivech povědomí nemá, webová archivní data vnímají jako nespolehlivá, a přestože si uvědomují význam dat pro další výzkum, neví, jak by je mohli využít a co by měli od webových archivů požadovat (Kvasnica, Rudišinová & Kreibich, 2016). Pokusili jsme se prozkoumat, jak se potřeby současných uživatelů webových archivů proměnily. Rozhodli jsme se zjistit formou dotazníkového šetření, jaké je povědomí o činnosti webového archivu v českém prostředí, jaké jsou potřeby a požadavky uživatelů na využití archivních dat.

Přípravě dotazníku předcházela rešerše zdrojů zaměřených na průzkumy uživatelských potřeb v zahraničí (Brejchová, 2021). Je z ní patrné, že se webové archivy napříč mezinárodní komunitou potýkají s podobnými problémy. Jednak je spojují srovnatelné procesní a technické parametry, v rámci evropského prostředí pak i obdobné právní podmínky, byť s drobnými nuancemi v rámci národních legislativ. Přestože povědomí o webových archivech i možnostech jejich využití roste, velká část akademiků o nich stále neví. Netuší, jaká data mají, nebo jaká by je mohla zajímat. Možnosti využití dat určuje legislativa omezující dostupnost zdrojů i nedokonalé nebo uživatelsky nevstřícné nástroje, chybějící vědecká infrastruktura, technické překážky v procesu archivace dat, potíže s interpretací dat vzhledem k jejich potenciální nespolehlivosti a neúplnosti (Costea, 2018). Jedním ze zmiňovaných projektů je i probíhající výzkum Jessicy Ogden, která zkoumá společenský význam webových archivů. Pokusila se charakterizovat několik typů účelu užití: citování (trvalý odkaz k obsahu), obcházení (paywally, dezinformace), monitoring (změny obsahu na webu nebo v chování zainteresovaných aktérů), vytěžování dat (strojové učení, trénování umělé inteligence) a zábava. Napříč rešeršovanými zdroji je zdůrazňována potřeba spolupráce kurátorů s vědeckou komunitou jak pro využití archivních

dat, tak i pro precizování akvizičních postupů. Vědci poukazují na nutnost technické dokumentace pro pochopení parametrů archivovaných dat (Kvasnica, Prokopová, Kvašová & Vozár, 2019), která mnohdy není dostačující nebo srozumitelná. Poznatky zahraničního průzkumu konvenovaly naší představě o uživatelských potřebách, přičemž k základním předpokladům patřilo, že zásadní překážkou pro práci s Webarchivem patří legislativně omezená dostupnost archivních dat. Naše hypotéza, na základě které byly rozpracovány výzkumné otázky, byla, že vědci povědomí o webových archivech i zájem o webová data mají, ale příliš je nevyužívají - nejen kvůli omezené dostupnosti, ale také kvůli nedostatečné informovanosti o archivních datech a o možnostech práce s nimi. Očekávali jsme, že respondenti budou mít podnětné náměty pro možné výzkumy podle specifik svého zaměření, ale nebudou vědět, jak s daty pracovat.

PRŮZKUM UŽIVATELSKÝCH POTŘEB

Příprava průzkumu probíhala v několika etapách. V první byla zpracována zmiňovaná rešerše zaměřená na potřeby uživatelů webových archivů v zahraničí (Brejchová, 2021). Byla zformulována hypotéza a vymezena cílová skupina respondentů. Záměrem průzkumu byla potřeba získat empirická data vztahující se k následujícím oblastem: Kdo jsou uživatelé webového archivu? Jaké mají informace o existenci webových archivů? Jaká je jejich zkušenost s použitím českého webového archivu? Za jakým účelem a jakým způsobem webový archiv používají / chtěli by používat? Jaké datové formáty, metadata, software využívají / chtěli by používat? Naší snahou bylo zapojit co nejvíce respondentů zejména z akademického prostředí. Pro průzkum jsme proto zvolili metodu kvantitativního výzkumu, která umožňuje oslovit velký okruh dotazovaných. Sestavili jsme strukturovaný dotazník, který neklade velké časové nároky na jeho vyplnění. Dotazník jsme se rozhodli primárně směřovat k badatelům a vědcům se zájmem o pokročilejší práci s našimi daty, než mají uživatelé individuální nebo institucionální.

Dotazník byl realizován v online podobě, zvolili jsme uživatelsky přívětivý Google Formulář, který umožňuje odpovědi jednoduše statisticky vyhodnocovat a graficky zpracovat. Struktura dotazníku byla rozvržena do čtyř oblastí. Obsahoval celkem 30 dotazů, kombinoval otázky uzavřené i otevřené. Základní dotazy byly povinné a bez jejich vyplnění nebylo možné pokračovat. U uzavřených otázek respondenti volili z nabízených možností, u otevřených otázek odpovědi nebyly standardizovány, formulovali je vlastními slovy. Zatímco uzavřené otázky jsou lépe kvantifikovatelné a lze je prezentovat prostřednictvím názorných vizualizací, otevřené otázky nabízejí prostor pro názory a nápady respondentů, které jsou pro nás velmi cenné. Každou oblast tvořil soubor několika dotazů. První část obsahovala identifikační otázky, zaměřené na zjištění obecných informací o respondentech. Následoval segment analytických otázek zařazených do sekce pojmenované Webarchiv. V této části jsme se zaměřili na získání informací o tom, jak se tazatelé orientují v problematice archivace webu. Závěrečné dvě části nazvané Funkcionalita a Data pak zahrnovaly dotazy, pro jejichž zodpovězení se předpokládá jistá zkušenost s archivací webu. Strukturu dotazníku jsme konzultovali s kolegy ze Sociologického ústavu AV ČR, v rámci pilotního průzkumu jsme dotazník předložili několika kolegům z akademického prostředí s žádostí o zpětnou vazbu. Náměty jsme do formuláře zapracovali, odpovědi z pilotního průzkumu jsem do výsledného hodnocení nezahrnuli.

Našemu úmyslu zaměřit dotazník zejména na vědeckou sféru jsme se snažili přizpůsobit způsob jeho distribuce k respondentům. S prosbou o šíření dotazníku jsme oslovili univerzity a vysoké školy, veřejné výzkumné instituce včetně jednotlivých pracovišť Akademie věd, krajské a vědecké knihovny a členské knihovny Asociace vysokoškolských knihoven. Pro šíření dotazníku jsme využili i mailové konference zahrnující kontakty širokého okruhu knihoven a snažili jsme se ho rozšiřovat prostřednictvím sociálních sítí (Twitter, Facebook). Naší snahou bylo získat co nejširší okruh respondentů z badatelské sféry, žádná

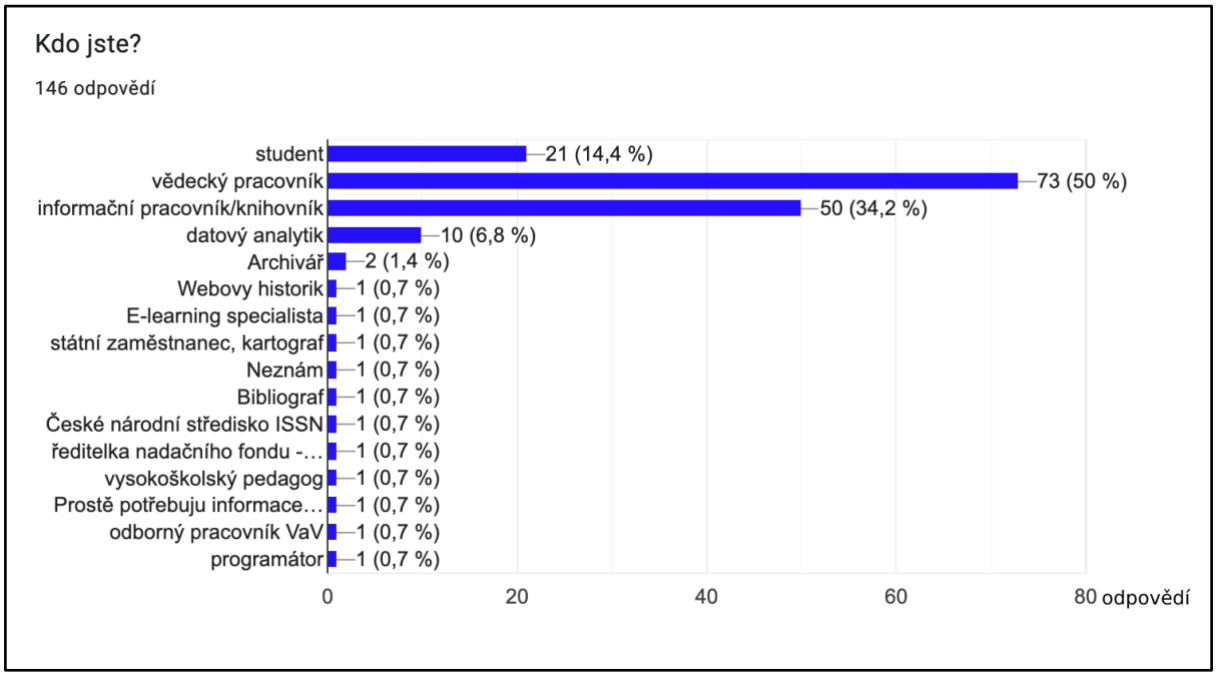
další kritéria pro vymezení vzorku nebo parametry definující jeho reprezentativnost jsme nestanovili. Zpracovali jsme všechny došlé odpovědi – celkem 146. Co jsme zjistili o potřebách uživatelů českého webového archivu?

VÝSLEDKY DOTAZNÍKOVÉHO ŠETŘENÍ

Kdo jsou uživatelé webových archivů

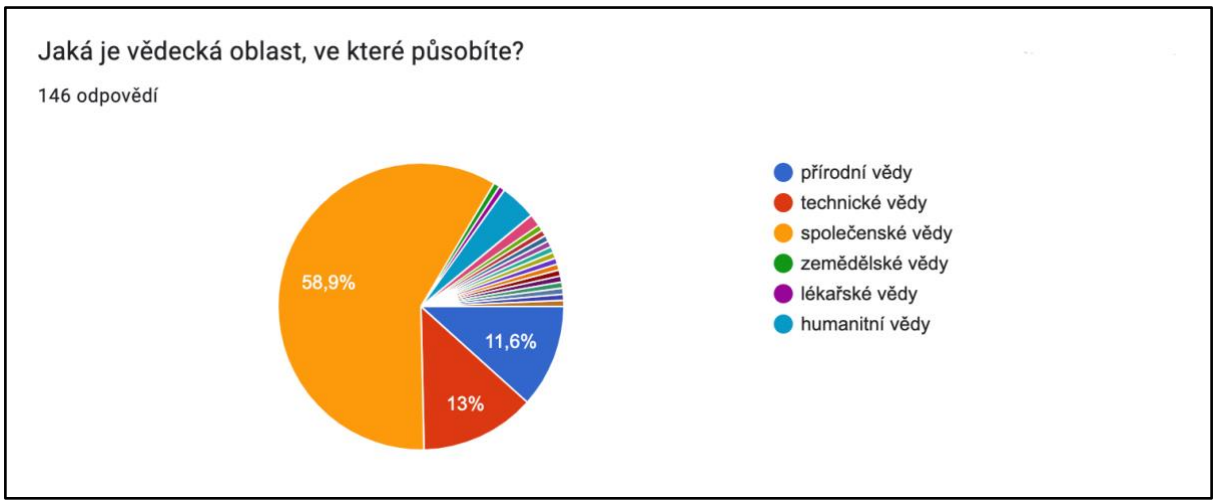
V první sekci otázek dotazníku jsme se snažili u respondenta identifikovat zejména to, z jakého oboru pochází, jaká je jeho pracovní pozice, v jaké instituci působí a jakému oboru se věnuje.

Jako vědecký pracovník se identifikuje 50 % respondentů. Druhou největší skupinou (34,2 %) oslovených je informační pracovník/ knihovník. Třetí nejpočetnější skupinou s 14,4 % jsou studenti. První dvě skupiny se mohou částečně prolínat, avšak vzhledem k formě dotazníku, kdy respondent vybíral z přednastavených odpovědí předpokládejme, že v případě „vědeckého pracovníka“ se o „informačního pracovníka/ knihovníka“ nejedná.



Graf 1 Kdo jste?

Co se týče oblasti působnosti, 58,9 % respondentů identifikuje svůj obor se společenskými vědami, 13 % s technickými a 11,6 % s přírodními.

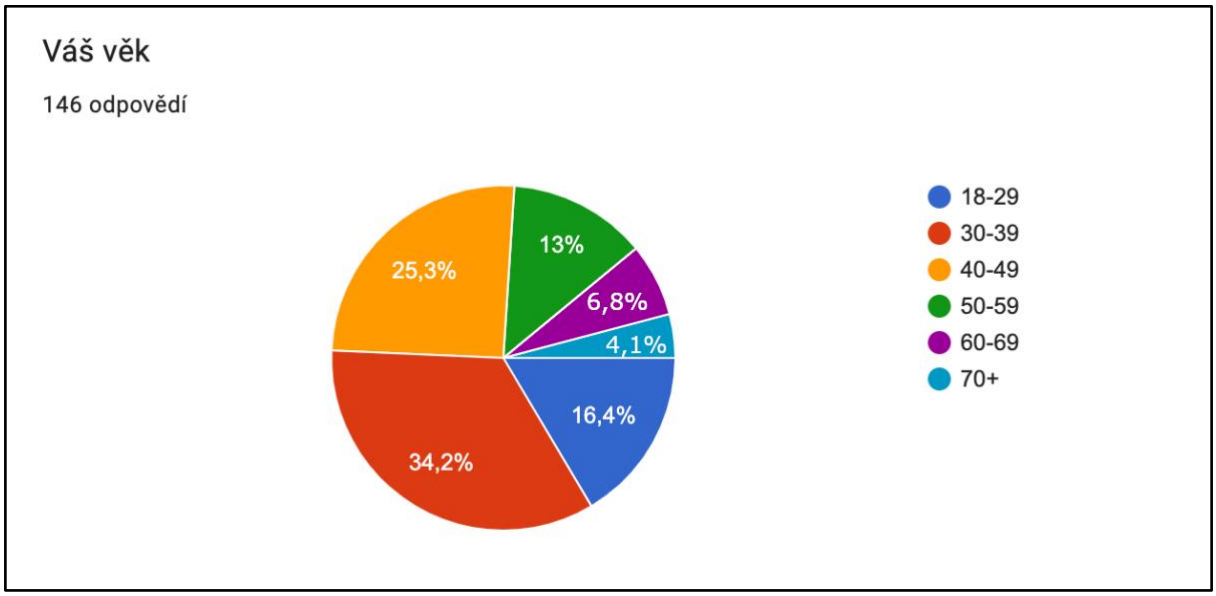


Graf 2 Jaká je vědecká oblast, ve které působíte?

V otázce, kde respondenti dostali za úkol popsat svůj obor, se vyhodnocení neobešlo bez čištění dat – odpovědi jsme redukovali na užší obory. Z výsledků vyšla jako nejpočetnější skupina zabývající se jazykovědou a literaturou 26 ze 146 (17,8 %), dále knihovnictvím 21 ze 146 (14,4 %), třetí byli informatici 8 ze 146 (5,5 %).

Největší počet respondentů se identifikuje s institucí Univerzity Karlovy – 26,7 % (z toho převládá FF UK – 15,8 % z celku), dále s Akademií věd ČR – 15,8 %. Zbytek jsou jednotky respondentů.

Z rozdělení respondentů z pohledu věkového vymezení vyplývá, že nejpočetnější skupinou (34,2 %) jsou uživatelé 30–39 let, dále 25,3 % 40–49 a 16,4 % tvoří respondenti ve věku 18–29 let.

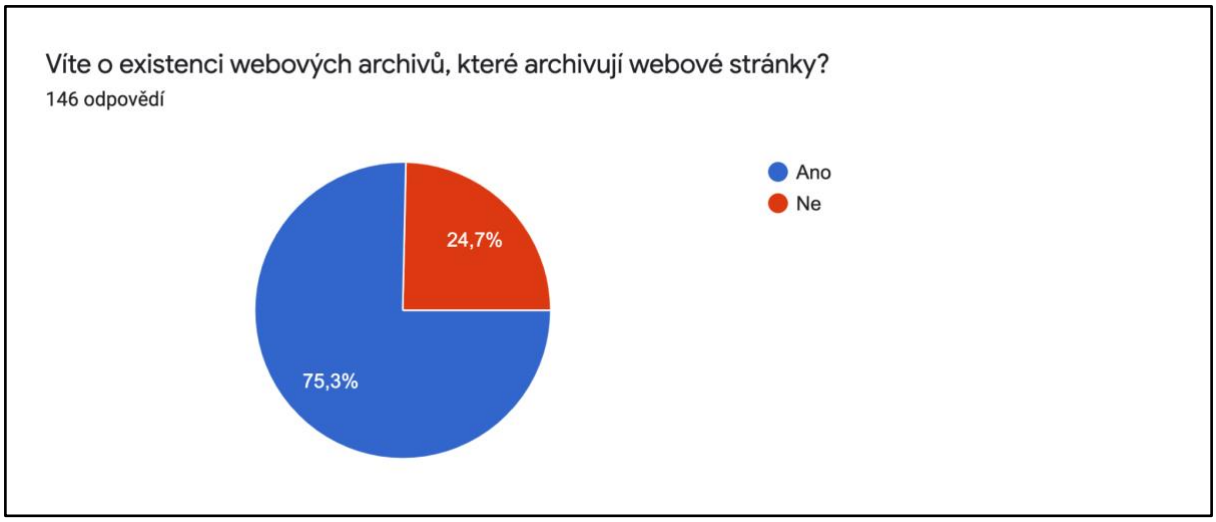


Graf 3 Váš věk

Nejpočetněji zastoupeným respondentem je tedy pracovník oboru společenských věd ve věku 30–39 let identifikující se s institucí Univerzita Karlova a zabývající se jazykovědou a literaturou.

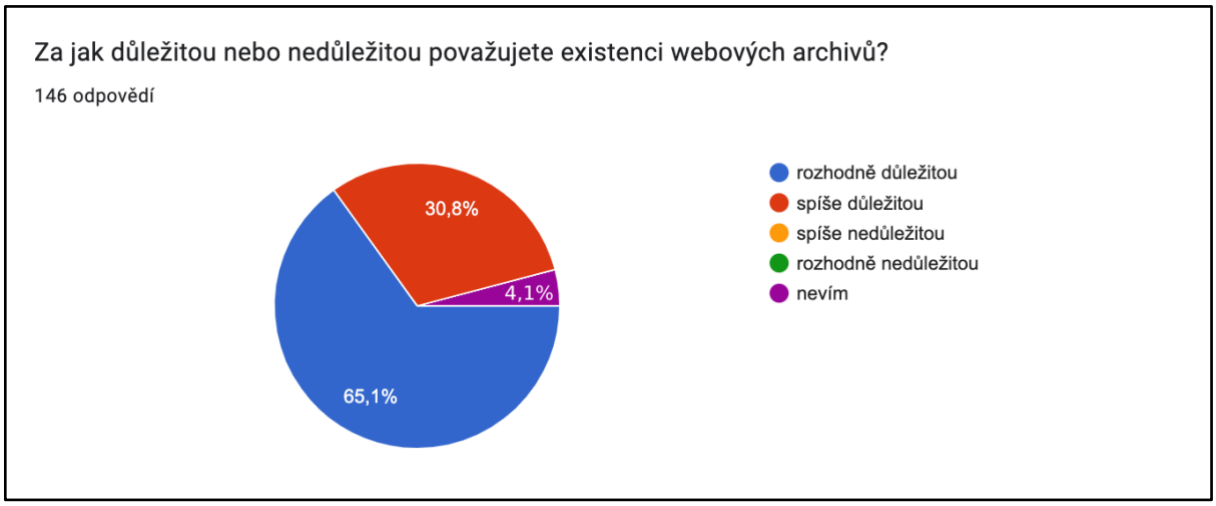
Povědomí o webových archivech

Většina (75,3 %) dotazovaných věděla o existenci webových archivů, zbylých 24,7 % nikoliv.



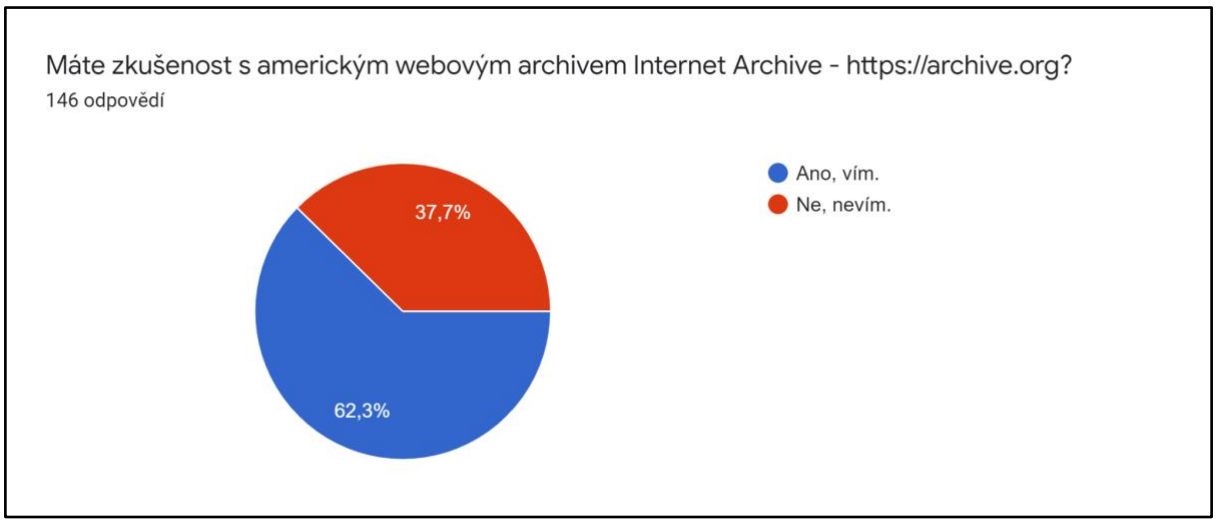
Graf 4 Víte o existenci webových archivů, které archivují webové stránky?

Více než polovina dotazovaných považuje existenci webových archivů za rozhodně důležitou – 65,1 % a 30,8 % respondentů vnímá webové archivy za spíše důležité.



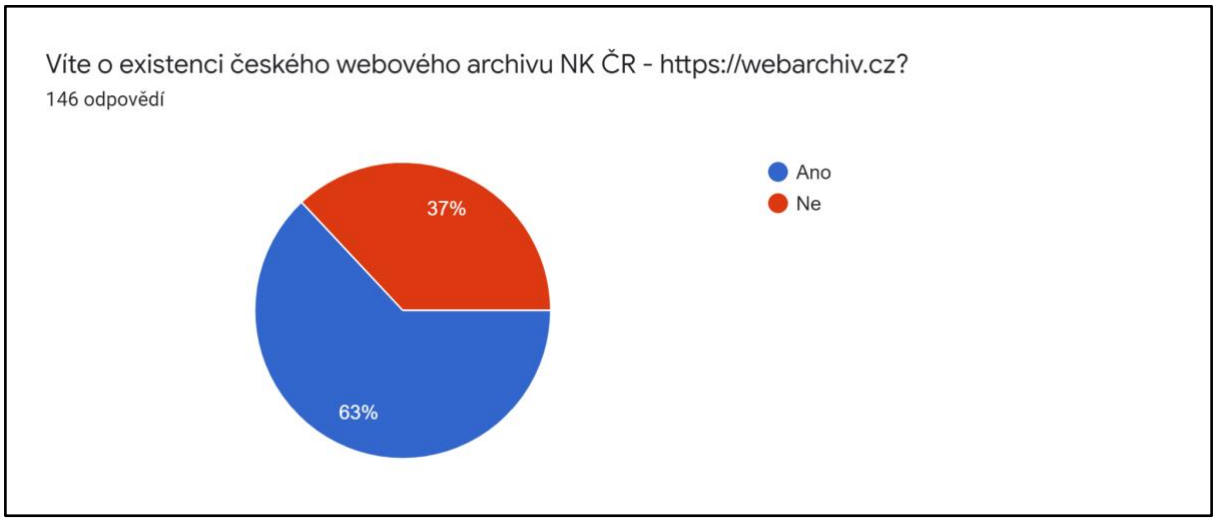
Graf 5 Za jak důležitou nebo nedůležitou považujete existenci webových archivů?

Internet Archive, službu, která je narozdíl od Webarchivu veřejně dostupná a obsahuje velké množství českých stránek, zná většina (62,3 %) oslovených.



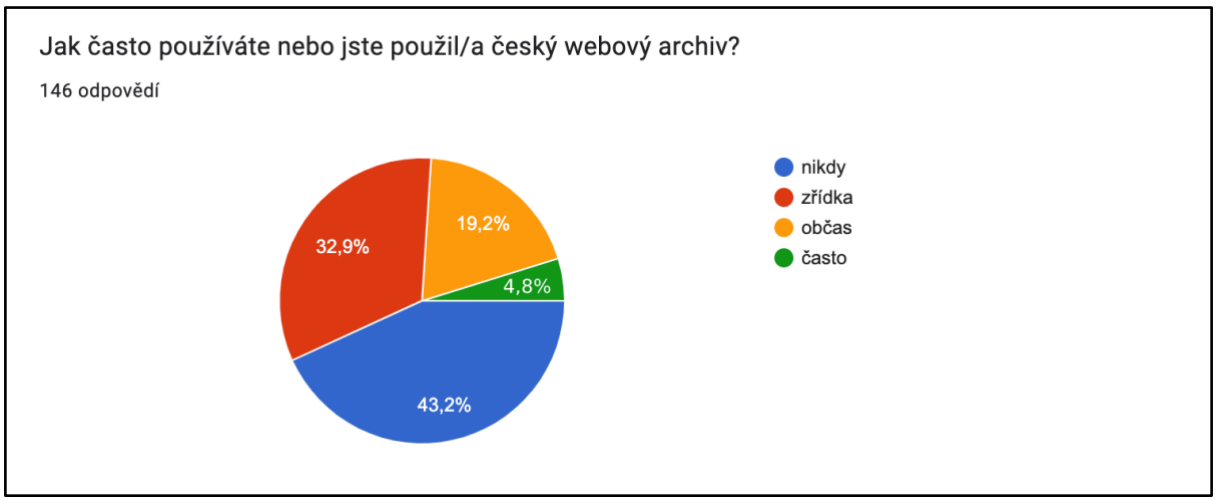
Graf 6 Máte zkušenost s americkým webovým archivem Internet Archive – https://archive.org?

O existenci českého webového archivu mělo povědomí 63 % oslovených a 37 % slyšelo o Webarchivu poprvé, poměr je tedy podobný jako u předchozí otázky.



Graf 7 Víte o existenci českého webového archivu NK ČR – https://webarchiv.cz?

Méně než polovina oslovených (43,2 %) odpověděla, že nikdy nepoužila Webarchiv, 32,9 % respondentů zvolila možnost zřídka, 19,2 % občas a 4,8 % respondentů Webarchiv používá často.



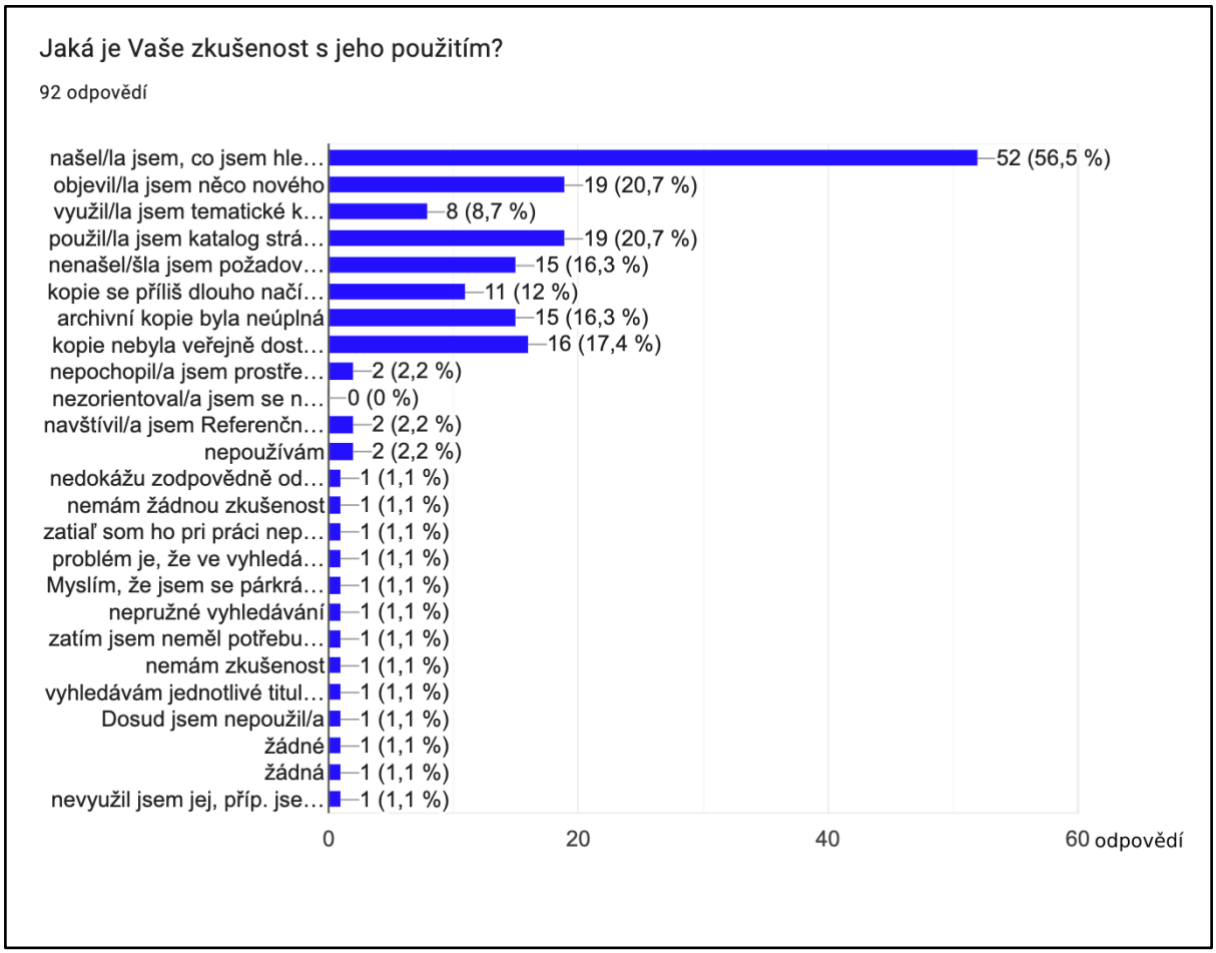
Graf 8 Jak často používáte nebo jste použil/a český webový archiv?

Na otázku „Pokud český Webarchiv znáte, jak jste se o jeho existenci dozvěděli?“ odpovědělo 85 respondentů, z nichž 16,5 % uvedlo, že Webarchiv zná ze studijního prostředí. O odpověď méně bylo první setkání v pracovním prostředí – 15,3 %. Přes grafický certifikát umístěný na webových stránkách vydavatelů zasmluvněných zdrojů se o Webarchivu dozvědělo 10,6 % z oslovených.

Zkušenosti s použitím Webarchivu

Většina reakcí byla pozitivní. Přesto jsme si vědomi, že služba má své slabé stránky – dostupnost webového archivu je pouze z Referenčního centra NK ČR, některé archivní kopie se načítají nekvalitně, mohou být neúplné apod. V reakci na otázku „Jaká je Vaše zkušenost s použitím Webarchivu?“ bylo možné volit z námi nabízených možností, případně je doplnit o vlastní (našel/la jsem, co jsem hledal/la; objevil/la jsem něco nového; využil/la jsem tematické kolekce webových zdrojů; použil/la jsem katalog stránek – webových zdrojů; nenašel/šla jsem požadovanou kopii; kopie se příliš dlouho načítala; archivní kopie byla neúplná; kopie nebyla veřejně dostupná; nepochopil/a jsem prostředí Webarchivu; nezorientoval/a jsem se na časové ose; navštívil/a jsem Referenční centrum NK ČR v Klementinu; jiná).

Více než polovina (56,5 %) z 92 dotazovaných zvolilo možnost, že našli, co hledali. 20,7 % respondentů objevilo něco nového, 20,7 % respondentů také použilo katalog webových stránek. 17,4 % respondentů uvedlo, že kopie nebyla veřejně dostupná a 16,3 % dotazovaných zvolilo, že archivní kopie byla neúplná.



Graf 9 Jaká je Vaše zkušenost s jeho použitím?

Otevřené otázky ponechávají svobodu ve formulaci odpovědi na rozhodnutí respondenta. Rozmanitost odpovědí sice předpokládá obtížnější zpracování či analýzu, na druhou stranu mohou být velkou inspirací pro směřování efektivnosti nabízených služeb Webarchivu. Z tohoto důvodu jsme se rozhodli zařadit do výzkumu několik otevřených otázek, odpovědi publikujeme v původní, neupravené podobě.

Na dotaz „S jakým konkrétním záměrem jste chtěl/a webový archiv použít?“ jsme získali následující odpovědi:

- *Starší programové prohlášení různých politických stran jazyková analýza*
- *Pro projekční činnost v železniční dopravě. Je často žádoucí vycházet a inspirovat se z historických údajů, které jsou hůře dohledatelné. Stará jízdní řády, dopravní provozu, stavební rekonstrukce tratí atd.*
- *Při vyhledávání např. názvu webových stránek či názvů/vyobrazení výrobků na nich (při zjišťování práv k užívání práv na označení – OZ, PVZ aj.)*
- *dohledávání dat Unie filmových distributorů, chtěla jsem zjistit, jestli je archivován web kritika Kamila Fily (Ještě větší kritik, než jsme doufali), což není tématicky zaměřený korpus*
- *např. všechny dokumenty zmiňující spartakiádu*

- *Už nevím – vím, že jsem hledala virtuální výstavy na archive.org, protože se jednalo o zahraniční výstavy a poté jsem citovala tyto projekty ve své práci s tím, že jejich trvalost je na tomto webu z mého pohledu reálnější.*
- *Vyhledávání informací k Latinské Americe, vzájemné vztahy mezi LA a ČR.*
- *kontrola záznamů v db ISSN*
- *Archivované stránky veřejné správy*
- *Výzkum soudobých dějin*
- *Využití pro bibliografickou databázi. Doplňuji URL odkazy do excerpovaných záznamů, aby si uživatelé databáze mohli v případě potřeby daný článek pročíst, pokud by původní odkaz již nefungoval.*
- *Doplňování odkazu na archivní zdroje u excerpovaných textů.*
- *pro výzkum jazyka*

Další z otevřených otázek byla „Za jakých podmínek / jakým způsobem by mohla být služba Webarchivu ve Vašem výzkumu použitelná?“ Níže uvádíme neredigovaný výběr nejpodnětnějších odpovědí.

- *no pokud bude vypadat jak ten americky archive …*
- *Zabývám se mj. korpusovou analýzou diskurzu, která pracuje s velkými objemy textů. Ve chvíli, kdy zjistím, že bych potřeboval texty, které na webu byly, ale už nejsou, může se mi webarchiv hodit. Víc to promyšleno nemám.*
- *Používáme WebArchiv při kontrole zdrojů, u nichž již přestala být původní URL adresa funkční (např. pro zjištění data ukončení vydáván/aktualizací zdroje)*
- *Pro studium vývoje určitých stránek, dohledávání informací*
- *Často hledáme konkrétní verze programů nebo analýz.*
- *Pro doplnění výukových materiálů, například jak se nějaká www služba během doby měnila a jaký to mělo dopad na obor.*
- *Dobrá a rychlá přístupnost Mapování vývoje (muzejních) webů, archivace virtuálních výstav nebo jiných online a pomíjivých výstupů digitálního kulturního dědictví.*
- *Hledání informací, které už najsou na webu dostupné.*
- *Zdroj informací, citace … úplné kolekce*
- *V momentě, kdy bude archiv přístupný z domu (bez nutnosti jet do Prahy do NK), 100% jej využiji*
- *vše veřejně dostupné, zlepšit vyhledávání*
- *keby bola voľne dostupná*

- *Jako trénovací data pro modely umělé inteligence*
- *ako iný druh prameňa dohledávání existujících či již vyčerpaných práv k duševnímu vlastnictví (nutné doložit verze webu vždy ke konkrétnímu datu).*
- *Ve svém výzkumu se zabývám českým internet a díky Webarchivu mám větší šanci, že linky na webové stránky, o kterých píši, sice nebudou za pár let funkční, ale k jejich obsahu se díky Webarchivu dostanu. Na některé weby ve svých textech odkazuji již přímo do Webarchivu, protože už neexistují. Díky Webarchivu lze také zpětně rekonstruovat tvář českého internetu.*
- *Vzhledem k zaměření spíše minimálně pro výzkum, ale velmi platná pro výuku.*
- *Pokud by její obsah byl mnohem kompletnější, např. na úrovni archive.org, který používám pravidelně kdyby byly výsledky dostupné*
- *Snad při vyhledávání starší odborné literatury? historické bádání v dějinách filmových institucí/ organizací, jejichž weby jsou archivovány, dohledávání recenzí publikovaných jen online, apod.*
- *Pokud by fungovalo vyhledávání pomocí hashtagů a jejich kombinací, případně by bylo možné hledat alespoň omezeně pomocí fulltextu.*
- *Případně by bylo fajn mít možnost vyhledávat podle více deskriptorů (jako v knihovním katalogu).*
- *Propojení s Alephem a jeho možnostmi by bylo super.*
- *Pokud bych se začala zabývat myšlením o literatuře posledních 20 let.*
- *Rozšíření přístupnosti i mimo referenční centrum NK dohledávám staré "aktuality" dle potřeb uživatelů*
- *Pokud by bylo možné stáhnout nějaké datové soubory k analýze např. přes API prostředí. Použitelná je už teď, ale šikovnější by bylo, kdyby více zdrojů bylo volně. Není-li zdroj volně, sáhnu po Internetovém archivu, protože ten je na rozdíl od Referenčního centra NK online.*
- *kdyby byla otevřená podobně jako internet archive výzkum, ke kterému se vracím zpětně, ale citované zdroje už na internetu nejsou dostupné*
- *Využívání pro nalezení starších článků, např. při katalogizaci starších zdrojů, u kterých se zrušila či změnila www adresa.*
- *Teoreticky kdyby uchovával i soubory ke stažení, které byly na stránkách umístěné, což je asi kapacitně nereálné a praktická využitelnost se limitné blíží nule.*
- *Hodně hodnotných projektů v češtině má i jiné domény ne cz, třeba com, net, nebo org kdyby měl API s korpusovým prohledáváním a mohla jsem si stáhnout materiály podle textového vyhledávání/metadat. Metadata by měla být uživatelsky srozumitelně dokumentována.*
- *dlouhodobé ukládání webů a jejich úprav archivace je v nasem oboru velmi komplexni problem, v soucasnosti mimo jine reseno v ramci iniciativy EOSC*

- *Archivační data ze starých článku nebo reportáží, tiskovek, smazané programy stran apod Asi v momentě, kdy bych analyzovala obsah webových stránek, které jsou v současnou chvíli nedostupné (dezinformační weby třeba)*
- *Pokud bych se více zabýval soudobými dějinami.*
- *Při zpracování badatelských dotazů, v závislosti na potřebách badatelů.*
- *Musí být umožněno vyhledávat na časové ose (jako ve službě archive.org). Stránky musí být veřejné (bez embarga, bez hlášení typu "you don't have permission to access ... on this server".*

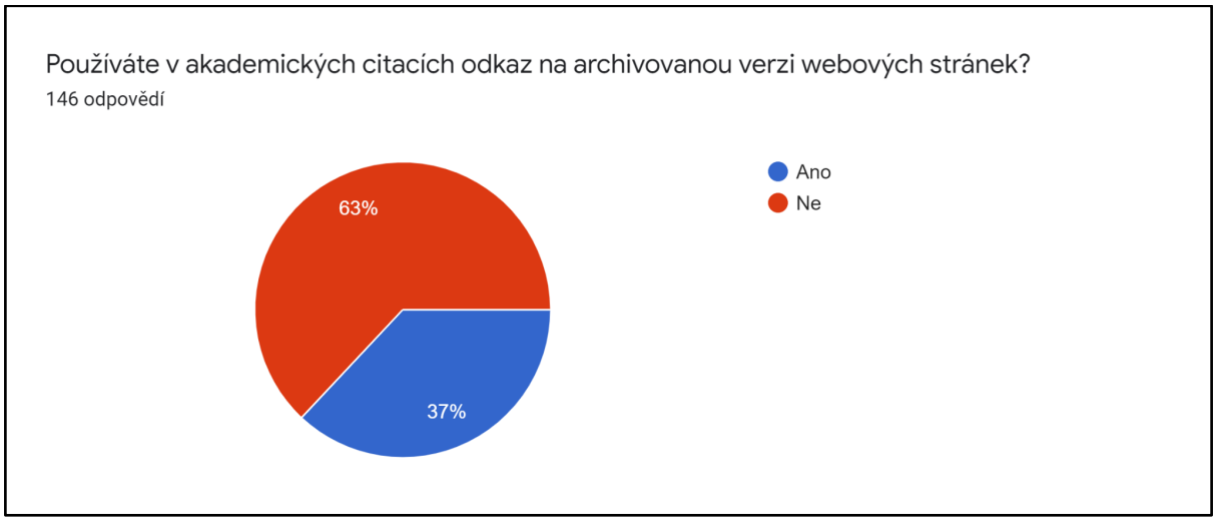
V 11 ze 73 odpovědí bylo zmíněno, že by respondenti službu použili, pokud by všechny zdroje byly zpřístupněné online.

Na otázku zaměřenou na možnosti případné spolupráce „Měl/a byste zájem se podílet na nominování webových zdrojů k dlouhodobé archivaci, či tvorbě tematických kolekcí a pomoci tak zachovat webové kulturní dědictví?“ odpověděla kladně menší část respondentů – 30,1 %. Návrhy webů k archivaci může zasílat kdokoli prostřednictvím formuláře na stránkách Webarchivu.

Graf 10 Měl/a byste zájem se podílet na nominování webových zdrojů k dlouhodobé archivaci, či tvorbě tematických kolekcí a pomoci tak zachovat webové kulturní dědictví?

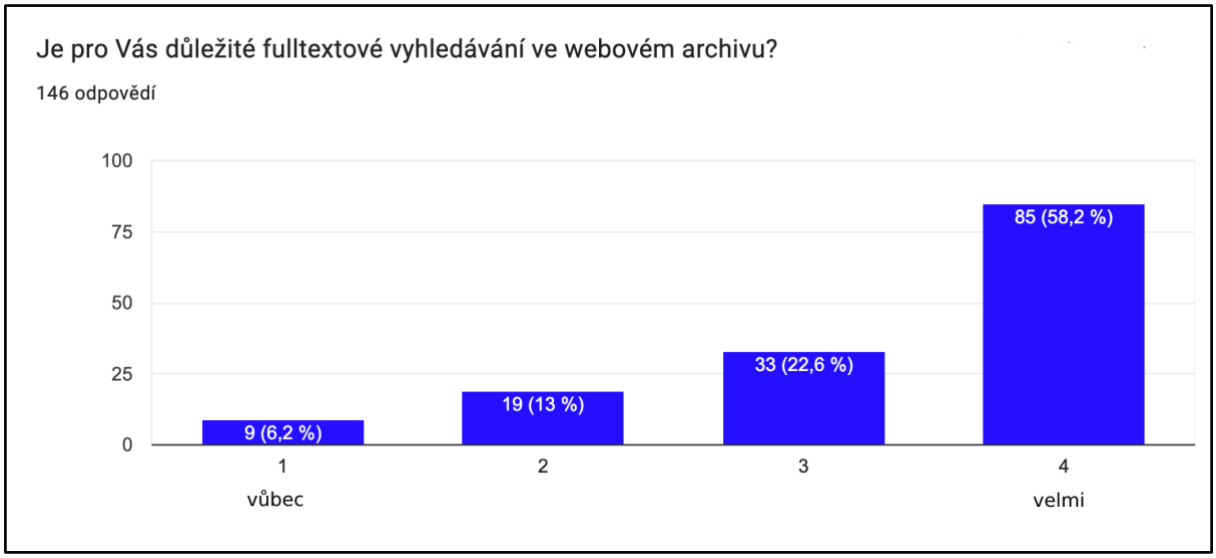
Funkcionalita

63 % respondentů nepoužívá v citacích v odborných pracích odkaz na archivovanou stránku v případě webových zdrojů.



Graf 11 Používáte v akademických citacích odkaz na archivovanou verzi webových stránek?

Význam přítomnosti fulltextového vyhledávání ohodnotilo jako *velmi důležité* 58,2 % respondentů. Jako *nijak důležité* 6,2 %.



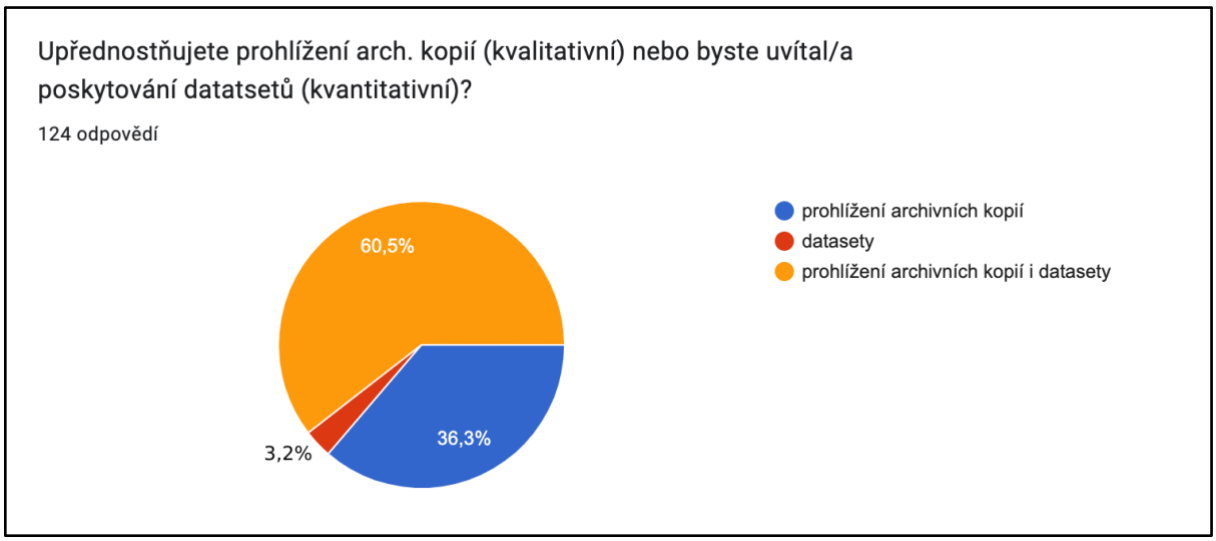
Graf 12 Je pro Vás důležité fulltextové vyhledávání ve webovém archivu?

Otázka po vyhledávacích parametrech vyjevila, že 44 odpovědí z 83 zmiňuje vyhledávání dle času (53 %), 13 dle URL (15,7 %), 18 (21,7 %) by rádo vyhledávalo dle textového vstupu, a 9 (10,8 %) dle tématu. U této otázky mohl každý respondent napsat více parametrů do jedné odpovědi.

Na otázku „Můžete uvést příklady výzkumů, v nichž si dokážete představit, že využijete či jste využil/a služby Webarchivu?“ bylo možné vyplnit cokoliv. Níže uvádíme neredigovaný výběr nejpodnětnějších odpovědí.

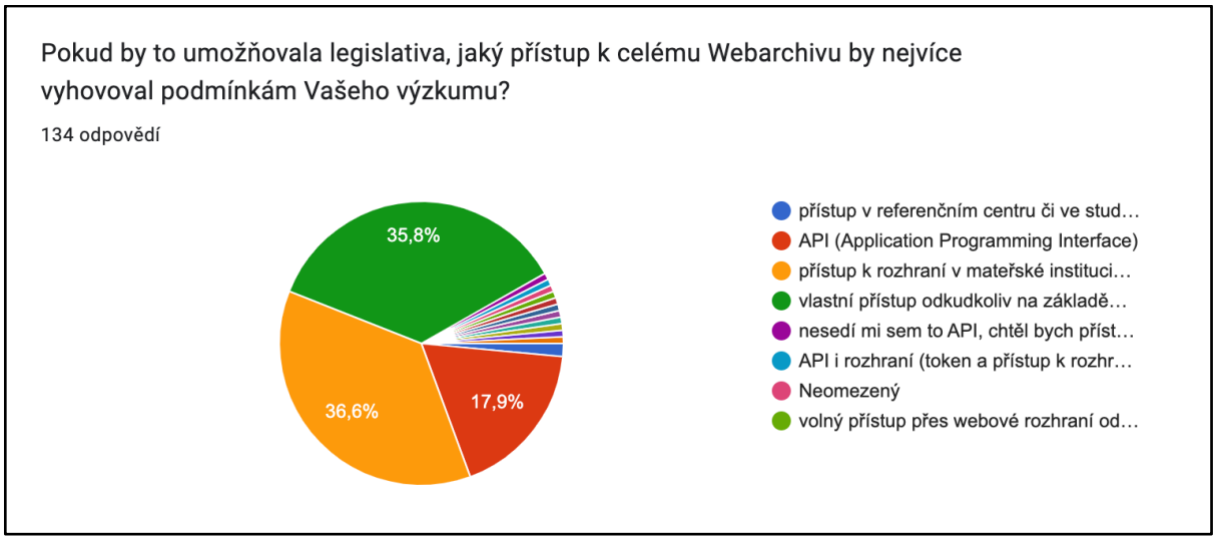
- *faunistika, lesnictvi – aplikovana entomologie, historie vyskumu, atd.*
- *vývoj webových stránek veřejné správy*
- *Projekce dopravních staveb a návrh jízdních řádů,*
- *monitorování GNSS incidentů*
- *Rešerše na podobnost práv na označení.*
- *Výzkum českého internetu – počty webů, typy webů, uživatelů, proměny komunikace na internetu, designu atd.*
- *retrokomparace*
- *Taxonomické databáze*
- *výzkum starého periodického tisku*
- *zmanopování české duchovní alternativní scéne scény porevoluční době, z té doby už spousta tránek neexistuje*
- *výzkum proměn češtiny v posledních desetiletích (postup přijímání nových slov, zastarávání slov, užívání frazeologie, slovních spojení...)*
- *Ano, například by se mohlo jednat o vyhledávaní informací na původních webových stránkách institucí, které na novou webovou stránku nepřenesly veškerý obsah. To se hlavně týká třeba stahovatelného obsahu (různých usnesení, zápisů z jednání, metodik atp.)*
- *zpětný přehled grantů, historie řešení dané problematiky*
- *hledání pramenů v oblasti výtvarného umění*
- *Zkoumání konkrétních témat v české společnosti po roce 1989 (např. debaty o Mukařovském, o Pravidlech českého pravopisu apod.) nicméně vše záleží na to, zda je materiál v daném zdroji dostupný. Nepostupuji tak, že bych si řekla "to je zajímavý zdroj, podívám se, co v něm je" ale naopak, hledám zdroj, který má něco k mému tématu…*
- *Demokracie, politika a svoboda, jejich chápání, vývoj a vliv na poznání a rozvoj společností*
- *dezinformační weby*

Většina respondentů (60,5 %) upřednostňuje jak poskytnutí datasetů, tak možnost prohlížení archivních kopií. 36,3 % odpovídajících by si vystačila jen s prohlížením archivních kopií a 3,2 % pouze s datasety.



Graf 13 Upřednostňujete prohlížení arch. kopií (kvalitativní) nebo byste uvítal/a poskytování datasetů (kvantitativní)?

36,6 % dotázaných označilo jako ideální přístup k Webarchivu z rozhraní v mateřské instituci. 35,5 % by uvítalo přístup odkudkoliv na základě badatelské smlouvy a 17,9 % respondentů by nejraději přistupovalo přes API.



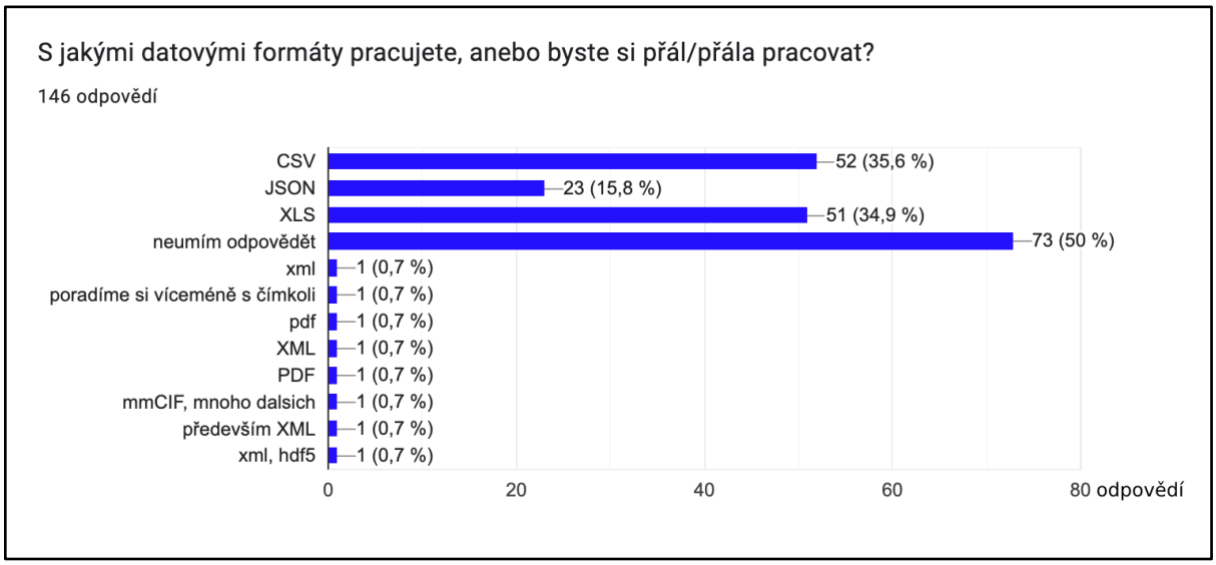
Graf 14 Pokud by to umožňovala legislativa, jaký přístup k celému Webarchivu by nejvíce vyhovoval podmínkám Vašeho výzkumu?

Nejpočetnější odpovědí na otázku týkající se preferovaných metadat pro výzkum se stala „klíčová slova“, která se v 73 odpovědích objevila 51krát. Další výrazně zastoupenou odpovědí byla „data sklízení“.

Práce s daty

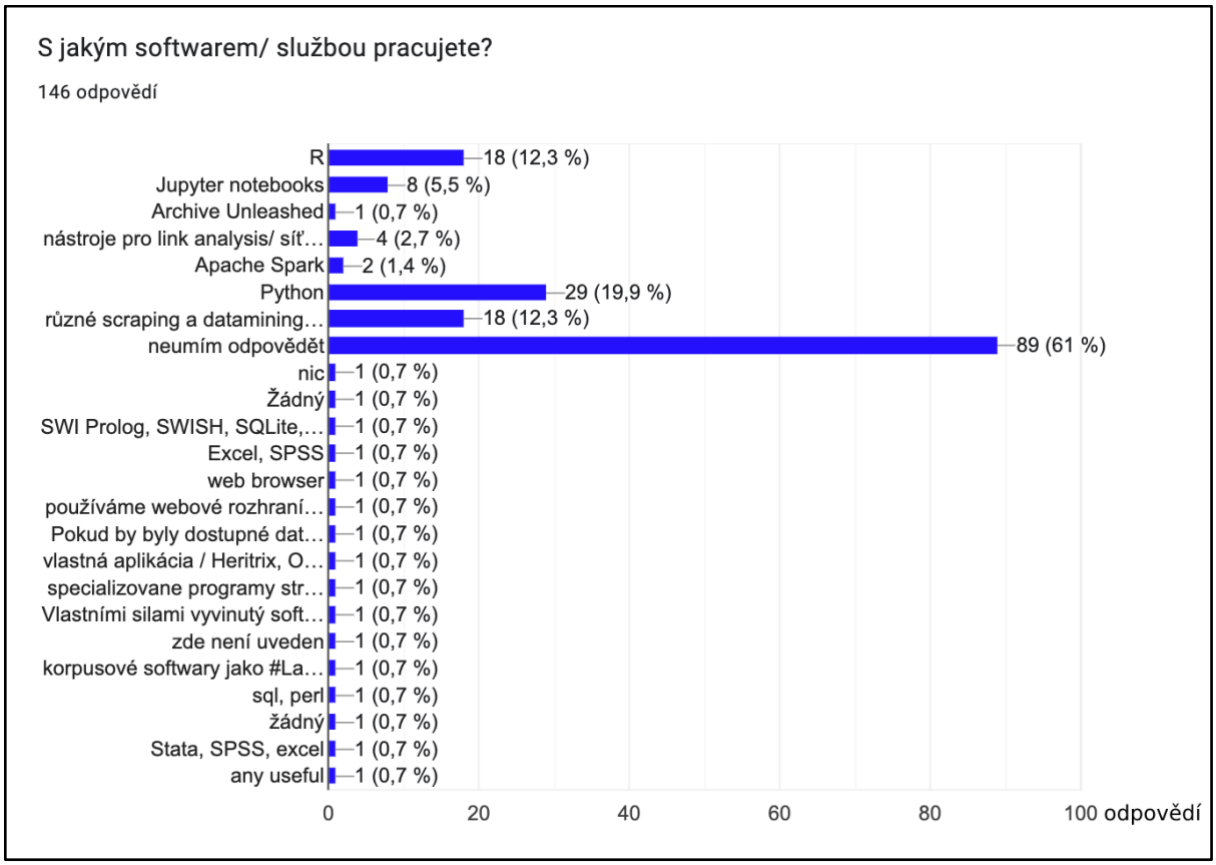
Na otázku „Jak objemný dataset a s jakými parametry byste si představoval/a? Jak nejlépe definovat dataset pro Vaši práci?“ padlo pouze 6 konkrétních odpovědí – respondenti tedy konkrétní představy o datasetu pro případný výzkum nemají.

Nejčastější konkrétní odpovědí na otázku „S jakými datovými formáty pracujete, anebo byste si přál/přála pracovat?“ se stal formát CSV (35,6 %), přestože 50 % respondentů vyplnilo, že „neumí odpovědět.“ Pro formát XLS se vyjádřilo 34,9 % dotazovaných.



Graf 15 S jakými datovými formáty pracujete, anebo byste si přál/přála pracovat?

Na otázku „S jakým softwarem/ službou pracujete?“ odpovědělo 61 %, že „neumí odpovědět“. 19,9 % uvedlo jako odpověď programovací jazyk Python. 12,3 % respondentů uvedlo software R, 12,3 % dotazovaných uvedlo nástroje pro sklízení webu a data mining a 5,5 % Jupyter Notebooks.



Graf 16 S jakým softwarem/ službou pracujete?

ZÁVĚR

Prezentované výsledky potvrdily v mnoha ohledech naše předpoklady. Badatelé povědomí o webovém archivu i zájem o jeho data mají, ale legislativní restrikce a z toho vyplývající omezený přístup k datům limitují jejich další využití. Přestože jsou v procesu legislativní změny alespoň ve smyslu otevření celého webového archivu v několika dalších knihovnách (tzv. Trojnovela) a zároveň probíhá implementace evropské směrnice o autorském právu na jednotném digitálním trhu do české legislativy, která se vztahuje k data miningu pro vědecké účely, v dohledné době nejsou další výrazné změny ve zpřístupnění dat očekávatelné. Většina dotazovaných považuje existenci webových archivů za důležitou a Webarchiv by používala, pokud by byly všechny kopie volně dostupné jako v případě amerického Internet Archive. Více než polovina dotazovaných uvedla, že zkušenost s Webarchivem má, převládala odpověď, že v archivu našli, co hledali. Byli jsme zvědaví na podněty uživatelů k možnostem využití dat a na nápady, s jakým záměrem nebo pro jaké výzkumy by badatelé chtěli nebo by si dokázali představit, že by bylo možné data použít. V textu uvádíme výčet těch z našeho pohledu nejpodnětnějších. Pokusili jsme se zjistit, jaké nástroje a formáty používají. Respondenti nemají většinou příliš představu o tom, jak by měl vypadat dataset pro jejich další výzkum, případně s jakým datovým formátem by chtěli pracovat nebo jaká metadata by uvítali. Zájemcům o data můžeme vyjít vstříc pokračováním ve vývoji uživatelsky přívětivých nástrojů, jako je zmiňované výzkumné rozhraní zaměřené na vytěžování velkých dat, které je možné používat bez nutnosti pokročilých IT znalostí. Pomocí kombinací různých funkcionalit a filtrů si mohou udělat představu o možnostech datasetu, který si mohou vytvořit na míru svým badatelským záměrům. I s přihlédnutím k výsledkům dotazníkového šetření plánujeme pokračovat ve snahách o zvyšování povědomí o archivaci webu prostřednictvím různých prezentací, workshopů, publikační činnosti – ať už odborné nebo třeba prostřednictvím komunikace na sociálních médiích. Popularizační efekt mělo i samotné šíření tohoto dotazníku. Za důležité považujeme v další naší činnosti i zpřístupnění dostatečných informací o parametrech archivních dat a o jejich metadatech a pokračování v rozvíjení spolupráce s vědci.

Výsledky dotazníku potvrdily i tušený stav ohledně četnosti používání archivované verze webu v citacích elektronických zdrojů. Více než polovina dotazovaných odkazy na archivní kopie v akademických pracích nepoužívá, což je škoda. Přestože Webarchiv umožňuje nominaci webového zdroje prostřednictvím formuláře, pro uživatele je zřejmě tato služba nedostačující, nemá o ní zájem, či neví o její existenci. Nabízí se do budoucna službu posunout blíže k uživateli, který by ji mohl použít přímo pro účel archivace citací, a tomu by nástroj měl uzpůsobit UIX design i funkcionalitu (vygenerování linku s archivovanou kopií *on demand*, což teď není možné). Nástroje pro archivace citací elektronických zdrojů by taktéž v druhém plánu mohly být cenným zdrojem pro tematické a výběrové sbírky obohacené o kontextová data zadaná samotnými uživateli. Uvést však tento nástroj v život tak, aby byl plošně používán, předpokládá další edukační, či normově integrační úsilí.

Webové archivy jsou unikátním zdrojem dat s velkým potenciálem pro další výzkum. Přestože archiváři pokračují v úsilí ve zpřístupnění dat a ve spolupráci s dalšími odborníky svůj zájem směřují k progresivním technologickým výzvám a tendencím, jako je využití umělé inteligence v praxi webových archivů (Svoboda, 2022), povědomí badatelů o možnostech užití archivovaných dat, jak vyplývá z našeho průzkumu, se mění pomalu. V rámci dotazníkového šetření jsme se pokusili zjistit, zda je Webarchiv využíván a kým, zmapovat povědomí o webovém archivu i nároky uživatelů na něj v českém prostředí. Ověřili jsme si některé naše hypotézy, konvenující zahraničním poznatkům, získali zpětnou vazbu a impulsy k zacílení našich výzkumných a popularizačních snah v následujících letech.

DEDIKACE

Studie a dotazníkové šetření byly realizovány na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace Národní knihovna České republiky poskytované Ministerstvem kultury ČR.

SEZNAM LITERATURY

Bailey, J. & Goel, V. (2016). *Program Models for Research Services.* University of North Texas Libraries, UNT Digital Library. https://digital.library.unt.edu/ark:/67531/metadc1477166/

Brejchová, I. (2021). *Uživatelé webových archivů a jejich potřeby: rešerše zdrojů a zahraniční literatury zaměřených na průzkum aktuálních uživatelských potřeb v oblasti elektronických, zejména webových zdrojů. Zhodnocení aktuálních poznatků.* Webarchiv, Národní knihovna ČR.

Costea, M. D. (2018). *Report on the Scholarly Use of Web Archives*. NetLab. http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf

ISO (2009). ISO 28500:2009 Information and documentation — WARC file format (2009). https://www.iso.org/standard/44717.html

Kvasnica, J., Rudišinová, B., Haškovcová, M., Holoubková, M., & Hrdličková, M. (2019). *Strategie budování sbírky Webarchivu: aktualizované znění.* Webarchiv, Národní knihovna ČR. https://webarchiv.cz/static/www/download/collection-policy.pdf

Kvasnica, J., Prokopová, A., Kvašová, Z., & Vozár, Z. (2019). Analýza českého webového archivu: provenience, autenticita a technické parametry. *ProInflow,* 11(1), 3–21. https://doi.org/10.5817/ProIn2019-1-2

Kvasnica, J., Rudišinová, B., & Kreibich, R. (2016). Vědecké využití dat z webových archivů. *Knihovna: knihovnická revue,* 27(2), 23–34. https://knihovnarevue.nkp.cz/archiv/dokumenty/2016-2/Kvasnica.pdf

Kvasnica, J., Vozár, Z., Haškovcová, M., & Kodad Holoubková, M. (2020). *Metodika pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu*. Národní knihovna ČR. http://invenio.nusl.cz/record/432325

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). https://apps.odok.cz/veklep-detail?pid=KORNBBXEMCL0

Proposal for a Directive of the European Parliament and the Council on copyright in the Digital Single Market COM/2016/0593 final - 2016/0280 (COD). (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593

Svoboda, L. (2021). Webarchiv spolupracoval na projektu Český literární internet. *E-zpravodaj Národní knihovny ČR,* 8(4), 6. https://www.nkp.cz/soubory/ostatni/ez_2021_4.pdf

Svoboda, L. (2022). Možnosti použití umělé inteligence pro webarchivační praxi. *E-zpravodaj Národní knihovny ČR,* 9(1)*, 9.* https://www.nkp.cz/soubory/ostatni/ez_2022_1.pdf

Vozár, Z., Haškovcová, M., & Prokopová, A. (2022). Internet jako pramen výzkumu: Přístup k archivovaným webovým zdrojům a možnosti jejich zpracování. *Teorie vědy/Theory of Science*. 1-29. https://doi.org/10.46938/tv.2022.552

Webarchiv. *Katalogizační manuál: katalogizační manuál pro popis elektronických online zdrojů ve formátu MARC 21.* https://webarchivcz.github.io/katalogizacni-manual/

Zákon č. 121/2000 Sb. Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon). (2000). https://aplikace.mvcr.cz/sbirka-zakonu/SearchResult.aspx?q=121/2000&typeLaw=zakon&what=Cislo_zakona_smlouvy

POZNÁMKA O AUTORECH

Marie Haškovcová

Marie Haškovcová je vedoucí Oddělení archivace webu Národní knihovny ČR, které spravuje Webarchiv – digitální knihovnu českých webových zdrojů. Je garantkou řešení oblasti Webové archivy pro vědecký výzkum v rámci institucionálního výzkumu NK ČR.

E-mail: marie.haskovcova@nkp.cz

Luboš Svoboda

Luboš Svoboda je kurátor – datový analytik webového archivu Národní knihovny ČR, digitální knihovny českých webových zdrojů. Je členem řešitelského týmu oblasti Webové archivy pro vědecký výzkum v rámci institucionálního výzkumu NK ČR.

E-mail: lubos.svoboda@nkp.cz

Markéta Hrdličková

Markéta Hrdličková je kurátorkou webového archivu Národní knihovny ČR, digitální knihovny českých webových zdrojů. Je členkou řešitelského týmu oblasti Webové archivy pro vědecký výzkum v rámci institucionálního výzkumu NK ČR.

E-mail: marketa.hrdlickova@nkp.cz