MOŽNOSTI ZPŘÍSTUPŇOVÁNÍ DAT Z WEBOVÝCH ARCHIVŮ

VÝVOJ CENTRALIZOVANÉHO ROZHRANÍ PRO VYTĚŽOVÁNÍ VELKÝCH DAT Z WEBOVÝCH ARCHIVŮ

Mgr. Marie Haškovcová, Bc. Andrea Prokopová

The article introduces a new phenomenon in the field of web archiving, specifically the efforts to work with large amounts of data and the use of metadata for accessing the archived data. Thanks to the growing demand for archived data, the project "Development of a Centralized Interface for Extracting Big Data from Web Archives" was created. This project aims to provide researchers with the data they need through a friendly user interface. In addition to the basic parameters of the project and its goals, the article describes the activities that accompany data archiving, such as the creation and storage of metadata, and the possibility of accessing archived data within the limits of the Czech legislation.

Keywords: web archiving; big data; data mining; metadata; social science research; software development

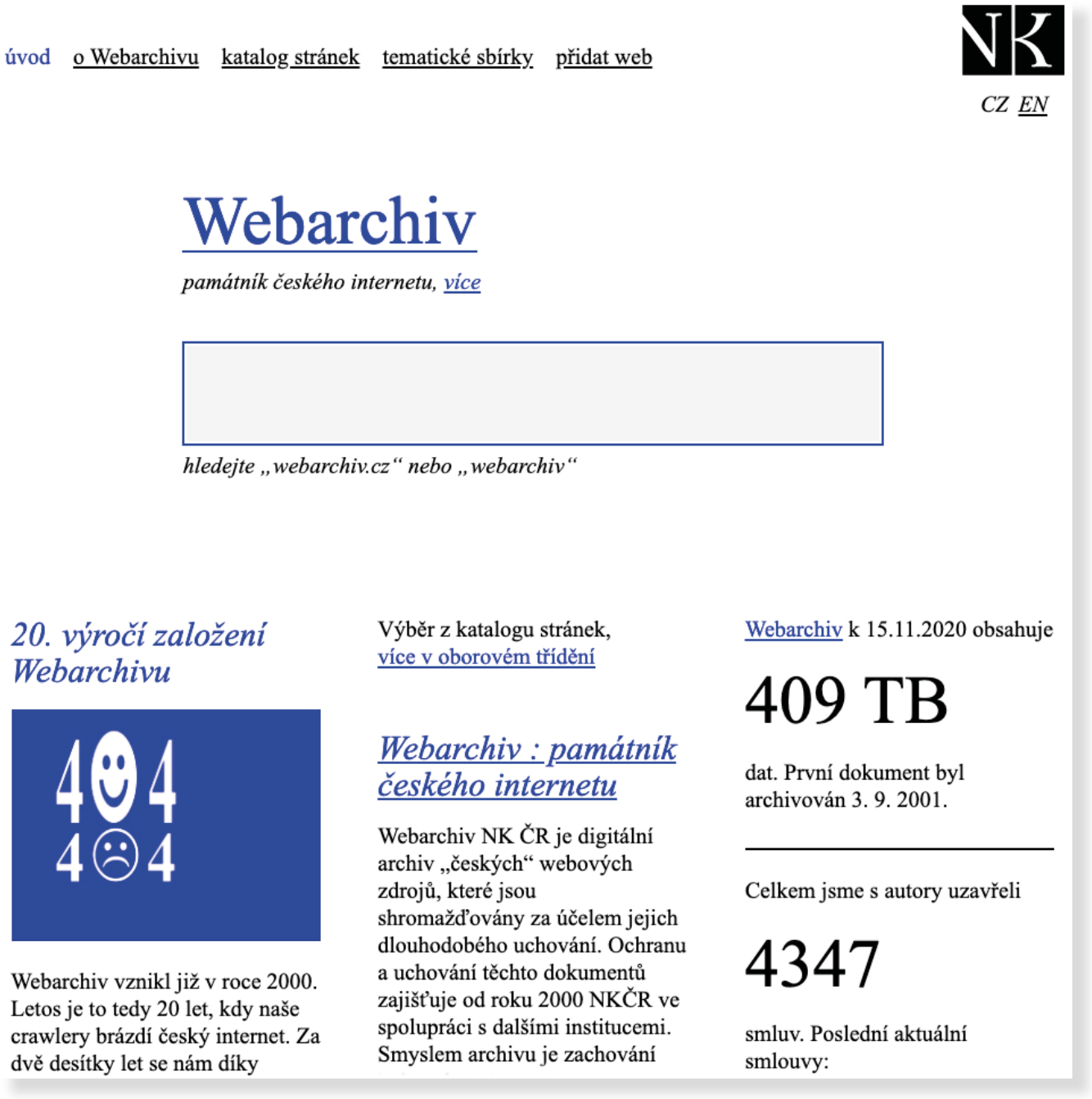
marie.haskovcova@nkp.cz
andrea.prokopova@nkp.cz

Národní knihovna
České republiky v Prahe
Oddělení archivace webu

https://www.nkp.cz/
https://webarchiv.cz

ARCHIVACE WEBU, MOŽNOSTI ZPŘÍSTUPŇOVÁNÍ DAT A WEBARCHIV

Internet je rozpínavé, extrémně proměnlivé a rychle se vyvíjející prostředí, probíhá v něm podstatná část lidské komunikace a interakce – veřejné i privátní. Data se rychle aktualizují, přesouvají, mizí. Paměťové instituce si proto na konci 90. let začaly klást otázky, jak online zdroje a born-digital dokumenty uchovat a ochránit. Postupně začala po celém světě vznikat pracoviště zaměřená na archivaci webu. K největším patří americký Internet Archive, který stál spolu s dalšími institucemi u vzniku profesní organizace IIPC – International Internet Preservation Consortium[1], kolem níž se sdružují instituce zaměřené na oblast archivace webu. Mezinárodní konsorcium podporuje sdílení zkušeností, formuluje tzv. best practices, vyvíjí open-source nástroje a snaží se poznatky formulovat tak, aby webové archivy mohly postupovat v dalším vývoji společně. V českém prostředí vzniklo v roce 2000 pracoviště zaměřené na archivaci webu v Národní knihovně ČR (dále NK ČR) – Webarchiv[2]. Patří k nejstarším archivům svého druhu a od roku 2007 je členem IIPC.



Webové rozhraní pro vyhledávání archivních kopií ve Webarchivu

Každý webový archiv si musí klást otázky, jak tento proměnlivý obsah zachycovat a dlouhodobě uchovávat, jak přistupovat k akvizici a jak obsah zpřístupnit svým uživatelům. Každé strategické rozhodnutí týkající se politiky sběru dat – z hlediska výběru obsahu i například parametrů nastavení sklízecích robotů – významně určuje jeho podobu. Možnosti zpřístupňování však výrazně limituje legislativa. Webarchiv se opírá o knihovní zákon, který mu umožňuje archivovat webový obsah pro archivní a konzervační potřeby a zpřístupnit obsah na půdě NK ČR. Přestože archivuje výhradně volně dostupný obsah, autorský zákon ukládá povinnost získat svolení se zpřístupněním archivních kopií na webu mimo budovu NK ČR. Webarchiv proto s vydavateli uzavírá licenční smlouvy, výjimku tvoří weby vystavené pod otevřenou licencí Creative Commons, která zveřejnění umožňuje. Uživatelům může proto volně zpřístupnit pouze 0,4 % z celého archivu. Vzhledem k velkému objemu archivních dat (Webarchiv disponuje 409 TB dat) a legislativním restrikcím hledá další způsoby, jak svůj unikátní fond zpřístupnit badatelské komunitě i široké veřejnosti. Zkoumá, jak lze zpřístupnit dostupná metadata, na která se autorský zákon nevztahuje, nebo jak nahlížet na archivní kopie jako na rozsáhlý datový set, nad nímž je možné provádět další analýzy. Jednou z takových cest je i projekt Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů[3]. V tomto ohledu bude klíčová i směrnice o autorském právu na jednotném digitálním trhu a její implementace do české legislativy[4], protože se zabývá výjimkami v oblasti data miningu pro instituce, které uchovávají kulturní dědictví.

ZPŮSOB UKLÁDÁNÍ DAT A TVORBA METADAT

Webarchiv používá open-source nástroje navržené archivářskou komunitou – pro sklízení robot Heritrix, pro zpřístupnění sklizených dat aplikaci Wayback Machine. Data v současnosti ukládá

39 lines (39 sloc) | 1.19 KB — Raw | Blame


```
{
  "_id": "5dcd98a695bcf5a1a194f0be",
  "recType": "harvest",
  "author": "NKCR",
  "date": "2019-11-14T19:10:46.983Z",
  "standard": "Grainery 0.4",
  "harvest": {
    "harvestPrefix:": {
      "harvestNameStand": "V6M_2017-10-05",
      "harvestFromWarcinfo": "V6M_2017-10-05",
      "harvestNameFNtrunc": "V6M_2017-10-05",
      "harvestDirsName": "V6M_2017-10-05",
      "harvestType": "?výběrová",
      "harvestSuffix": ["V6M", "2017-10-05"]
    },
    "date": "2017-10-05T11:26:00.000Z",
    "harvestID": "105f23c9-b037-4c1d-901c-dcf272877d9f",
    "size": 618029,
    "warcsNumber": 12092
  },
  "harvestCrawl": {
    "logs": true,
    "path": "logs/crawl",
    "fileName": ["crawler00.tar.gz", "crawler01.tar.gz", "crawler03.tar.gz"]
  },
  "paths": {
    "cdxsID": ["105f23c9-b037-4c1d-901c-dcf272874d9f"],
    "warcsID": ["105f23c9-b037-4c1d-901c-dcf272877d9f"],
    "warcsFilenames": ["V6M_2017-10-05-crawler00.webarchiv.cz-warc.gz"]
  },
  "revision": {
    "dateOfValidation": "2019-12-04T19:10:46.983Z",
    "statusOfValidation": "NA",
    "nextLastDateOfValidation": "2021-12-03T19:10:46.983Z",
    "hashOrig": "NA",
    "hashLast": "NA",
    "commentaries": { "exists": false, "text": "NA" }
  }
}
```


Příklad metadatového záznamu ve formátu JSON - sklizeň (harvest), administrativní a technická metadata

ve standardizovaném kontejnerovém formátu WARC[5]. V archivu se nacházejí i archivní balíčky v již zastaralém, nestandardizovaném formátu ARC, který WARCu předcházel. Jak vypadá metadatový popis sklizených dat, který umožňuje dále s nimi pracovat, vytvářet různé analýzy nebo například definovat množinu dat pro vědecký výzkum? Webarchiv pracuje s dvěma typy metadat – jednak s popisnými, jednak s technickými a administrativními. Popisná metadata jsou zpracovaná pouze k malé části zdrojů – těch, které jsou součástí výběrových sklizní. Zdroje, které jsou volně dostupné, mají zpracovaný podrobný katalogizační záznam, a to v knihovním systému Aleph, ve formátu MARC a od roku 2015 podle pravidel RDA. Jejich podrobný popis je k dispozici v Katalogizačním manuálu[6]. Takto pořízené záznamy jsou součástí České národní bibliografie, v katalogu NK ČR mají vlastní bázi nazvanou Elektronické zdroje. Druhým typem jsou metadata technická a administrativní. Obsahují podrobné informace o souborovém formátu, ve kterém jsou data uložena, a popisují technické údaje získané během sklizní. Vztahují se k sklizním, ke kontejnerovému formátu a k indexu[7]. Příkladem základních údajů může být datum zahájení a ukončení sklizně, její typ, rozsah nebo autor. V roce 2020 Webarchiv zpracoval Metodiku pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu, certifikovanou Ministerstvem kultury ČR[8]. Metodika předkládá postup pro strukturu metadatového záznamu a jeho popis, přičemž vychází ze softwaru Grainery[9], který lze použít pro generování, extrakci, evidenci a zobrazení metadat. Umožňuje tak s nimi lépe pracovat a zapojovat je do dalších výzkumů.

Webarchiv Katalogizační manuál

Katalogizační manuál pro popis elektronických online zdrojů ve formátu MARC 21.

Katalogizační manuál – Úvod / FMT / LDR / BAS / Kontrolní pole / Identifikační čísla a kódy / Hlavní záhlaví / Název a další názvové údaje, nakladatelské údaje / Údaje fyzického popisu / Poznámky / Věcné selekční údaje / Vedlejší záhlaví / Propojovací pole / Alternativní prezentace / Kódy zpracovatelské instituce /

Alternativní prezentace, elektronické umístění

856 (Elektronické umístění a přístup) (O)

Pokud je tentýž zdroj dostupný z více URL adres, další URL se zapisuje/-í vždy do nového pole 856.

Pokud je/jsou starší URL již neplatné a zapisuje se nová (tj. platná) URL, zapíše se tato nová URL do 1. výskytu, před původní a již neplatnou URL. Totéž platí, je-li původní URL přesměrována na nově zapisovanou URL (tj. nová URL = 1. výskyt, přesměrovaná URL = 2. a případně další výskyty).

Pro odkaz do Webarchivu se používá další výskyt pole 856, zápis je stejný jako u běžné adresy v poli 856, ale v podpoli $z je hodnota „archivní verze stránek“ (text, který se zobrazuje za URL adresou).

```
např.:
856 40 $u http://gvuhodonin.cz
       $q text/html
       $4 N
856 40 $u http://wayback.webarchiv.cz/wayback/gvuhodonin.cz
       $q text/html
       $z archivní verze stránek
       $4 N
```

Katalogizační manuál pro popis elektronických zdrojů ve formátu MARC 21

VÝVOJ CENTRALIZOVANÉHO ROZHRANÍ PRO VYTĚŽOVÁNÍ VELKÝCH DAT Z WEBOVÝCH ARCHIVŮ

Nastíněný legislativní rámec i metadatové specifikace se staly výchozími body pro uvažování o dalších možnostech využití archivních dat. Jedním z cílů aktuálně realizovaného výzkumného projektu Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů[10] je funkční a efektivní řešení, jak legálně zpřístupnit sklízená data vědecké obci. Potenciál dat, kterými Webarchiv disponuje, je v současné době využíván jen minimálně i kvůli české legislativě. Zájem vědců o archivovaná data roste. A tak se zvyšuje i poptávka po způsobech, jakými lze velké objemy dat zpracovat. Sílící zájem potvrzuje mimo jiné i první ročník mezinárodní konference EWA: Engaging with Web Archives, která se konala na podzim roku 2020. Zaměřila se na využití webových archivů v oblasti výzkumu a vzdělávání v řadě oborů a byl na ní představen i tento projekt[11]. Dá se proto očekávat, že se budou měnit i požadavky na současné infrastruktury, které už nebudou sloužit jen k archivaci, ale bude možné je využít i k analýze a další práci s datasety.

Na projektu spolupracují tři instituce. Vedle NK ČR je to Sociologický ústav Akademie věd ČR, v.v.i. (dále SOÚ), který pracuje s velkými objemy dat v sociálně vědní oblasti, a Fakulta aplikovaných ved Západočeské univerzity v Plzni (dále ZČU), která navrhuje technická řešení, k nimž patří například strojové zpracování velkých objemů dat nebo automatické rozpoznávání informací z video nebo audio souborů. Ve spolupráci se SOÚ, jehož výzkumné požadavky v rámci projektu definují potřeby badatelské komunity, se ZČU věnuje analýze témat textového dokumentu a jejich automatické detekci. Pro klasifikaci dokumentů používá nejaktuálnější přístupy založené na hlubokých neuronových sítích[12]. Softwarové řešení vzniká ve spolupráci NK ČR a ZČU s externí firmou InQool, a.s.

WORKFLOW A PŘÍNOS PROJEKTU

Projekt je rozdělen do tří etap. V rámci vstupní fáze proběhla hloubková analýza dat webového archivu[13]. Cílem bylo zjistit, jak sklízená data vypadají a která z nich bude možné využít k dalším výzkumům. V první fázi byly rovněž definovány otázky sociálně vědního výzkumu a vhodný metodologický přístup. Následující analytická fáze se zaměřila na extrakci vzorových datasetů a jejich analýzu, vytvoření strukturovaného indexu a vývoj nástrojů potřebných pro sémantickou analýzu textu. Cílem poslední implementační fáze projektu je vývoj centralizovaného rozhraní k vytěžování velkých objemů dat a vytvoření grafického rozhraní pro uživatele, které bude sloužit k vyhledávání výzkumných datasetů.

Kromě výzkumných cílů a nových technologických posunů a inovací se projekt věnuje i přezkoumání právního rámce. V rámci projektu byla vytvořena právní analýza, která popisuje postavení webového archivu NK ČR a možnosti využití dat s ohledem na autorské právo. V dalších fázích se bude věnovat aktuálním právním otázkám, k nimž bude patřit posouzení implementace výše zmiňované směrnice o autorském právu na jednotném digitálním trhu v českém prostředí nebo tzv. trojnovela[14], která mimo jiné zakotvuje právní úpravu web-harvestingu. V platnost by měla vstoupit v roce 2022. V technické části projektu aktuálně probíhá implementace nových řešení, k nimž patří rozhraní APACHE-HADOOP,[15] které slouží ke klastrování dat.[16] Pro práci s daty byl navržen tzv. intermediary formát, jenž slouží k očištění dat od sekundárních informací a usnadňuje tak následné vyhledávání a export požadovaných dat. Výstupem projektu bude fasetový a fulltextový vyhledávač pro práci s objemnými soubory dat, jehož součást bude tvořit i speciální exportní aplikace pro extrakci zvolených dat. Tento vyhledávač tak umožní vědcům vybrat si část dat z datasetů, která potřebují pro své analýzy, a zúročit tak ve zcela nových perspektivách bohatství archivovaných dat, která Webarchiv za dvacet let své existence nashromáždil.

POZNÁMKY

[1] https://netpreserve.org.

[2] https://www.webarchiv.cz.

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

[4] Plné znění směrnice: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593.

[5] https://www.iso.org/standard/44717.html.

[6] https://webarchivcz.github.io/katalogizacni-manual/.

[7] Index je databázová konstrukce sloužící ke zrychlení vyhledávacích a dotazovacích procesů v databázi. Představuje kompletní seznam archivních objektů v kontejneru a poté jejich umístění v celém archivu.

[8] http://invenio.nusl.cz/record/432325.

[9] https://github.com/WebarchivCZ/grainery, https://github.com/WebarchivCZ/grainery/wiki.

[10] Informace k projektu a výstupy: https://www.webarchiv.cz/vyvoj/.

[11] https://ewaconference.com, https://zenodo.org/record/4058013#.X622lJNKjeo.

[12] LEHEČKA, et al. LEHEČKA, Jan et al. Adjusting BERT's Pooling Layer for Large-Scale Multi-Label Text Classification. Text, Speech, and Dialogue: 23rd International Conference [online]. Brno, 2020 [cit. 2020-11-12]. ISBN 978-3-030-58323-1. ISSN 1611-3349. Dostupné z: https://doi.org/10.1007/978-3-030-58323-1.

[13] KVASNICA, Jaroslav, et al. Jaroslav et al. Analýza českého webového archivu: Provenience, autenticita a technické parametry. ProInflow: Časopis pro informační vědy [online]. 2019, 11(1), 3-21 [cit. 2019-11-20]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/2019-1-2.

[14] https://apps.odok.cz/veklep-detail?pid=KORNBBXEMCL0.

[15] https://hadoop.apache.org/

[16] Klastrování dat, neboli shluková analýza, je proces, kdy se vytvářejí kategorie na základě podobnosti a stejných vlastností objektů.