Internet archive
https://archive.org/

Andrej Betík, 4. ročník

20.11.2019

1 Charakteristika projektu

Internet archive je digitálna knižnica, ktorá predstavuje rozsiahle úložisko rôznych internetových dát. Táto knižnica poskytuje voľný prístup k archívnym webovým stránkam, knihám, audio nahrávkam, videám, obrázkom a softwarovým programom [1].

2 Doba riešenia projektu, aktuálny stav

Projekt Internet archive založil Brewster Kahle v roku 1996 ako neziskovú organizáciu 501(c)(3), ktorá mala spočiatku za úlohu výhradne archiváciu webových stránok. V roku 1999 sa archív rozšíril o kolekciu amerických historických filmov [2]. V roku 2012 začala organizácia Internet archive poskytovať možnosť sťahovania súborov pomocou služby BitTorrent [1], ktorá výrazne zrýchlila proces sťahovania. Aktuálne archív disponuje viac ako dvadsať ročnou históriou webu, ktorá zahrňuje viac ako tristotridsať biliónov webových stránok, dvadsať miliónov kníh a textov, päť miliónov videí a dvestotisíc softwarových programov. Projekt je financovaný prostredníctvom darov, grantov a poskytovaním archivačných služieb. Medzi nadáciami, ktoré tento projekt veľkoryso podporili patrí napríklad Knight Foundation alebo Andrew W. Mellon Foundation. Zoznam všetkých nadácií je dostupný na webovej stránke Internet archive [1].

3 Ciele projektu

Organizácia Internet archive sa snaží uchovať ľudské poznatky a kultúru vytvorením digitálnej internetovej knižnice dostupnej historikom, výskumníkom, študentom a širokej verejnosti [3]. Taktiež sa usiluje o spoluprácu s inými organizáciami pri súvisiacich projektoch, ako napríklad, bojovanie s dezinformáciami a falošnými správami. Jej hlavným a dlhodobým cieľom je poskytovať univerzálny prístup ku vedomostiam [1].

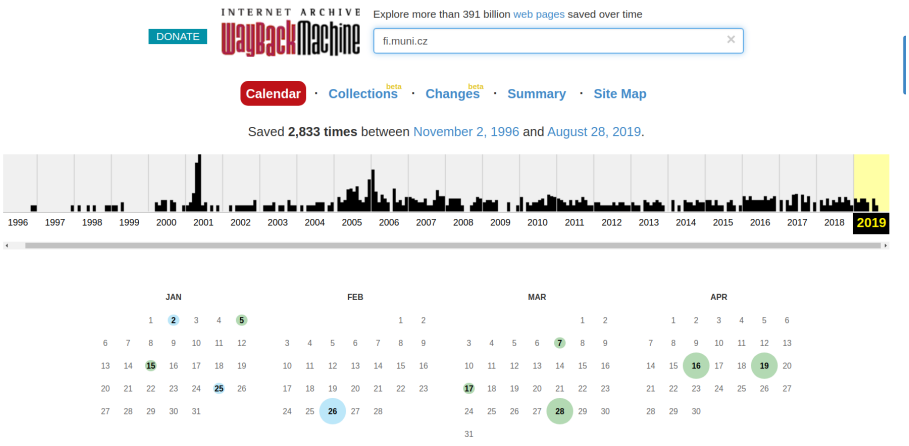
[1] BitTorrent je peer to peer internetový protokol, ktorý poskytuje rýchly, jednoduchý a efektívny prenos veľkých súborov medzi ľubovoľným počtom ľudí

4 Popis projektu a jeho výsledkov

Digitálna knižnica Internet archive poskytuje mnoho služieb, ktoré sa delia podľa formátu dát do nasledovných kategórii: archivácia webu, kníh a textov, videí, audio nahrávok, softwarových programov a obrázkov.

Archivácia webu

Wayback machine je služba, ktorá poskytuje jednoduché vyhladanie archívnych verzií webovej stránky zadaním adresy danej webovej stránky do vyhľadávača. Inšpirácia na jej vytvorenie plynie z problému zanikania obsahu stránok pri ich zmene alebo likvidácii. Služba zbiera webové stránky pomocou takzvaných "crawlerov." Tento prístup je založený na automatickom prehľadávaní a uchovávaní stránok. Toto automatické zbieranie webových stránok je obmedzené stránkami, ktoré sú neprístupné alebo sa na nich vzťahujú autorské práva. Wayback machine poskytuje prehľadnú analýzu archívu webovej stránky v podobe kalendáru stiahnutí danej webovej stránky (viz Obrázok 1).



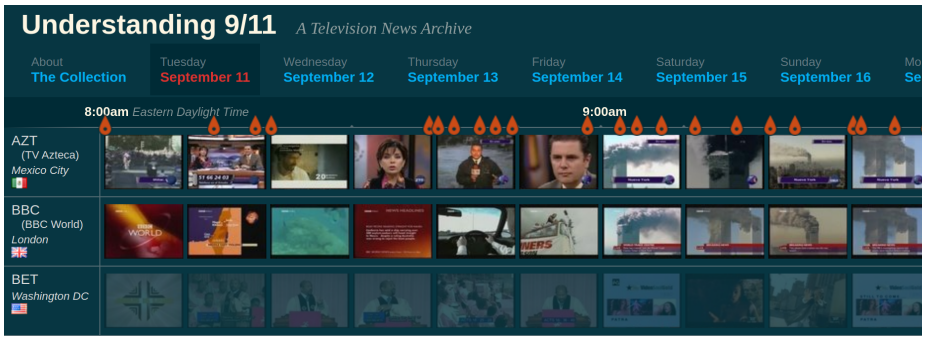
Obrázek 1: Vyhľadávanie archívych verzií webovej stránky pomocou Wayback machine

Archivácia kníh a textov

Keďže Internet archive je primárne knižnica, venuje osobitú pozornosť knihám a textom. Organizácia spolupracuje s viac ako tridsiatimi troma centrami, ktoré skenujú vyše tisíc kníh každý deň. Aktuálne je k dispozícii viac ako dvadsať miliónov kníh a textov [4]. Internet archive tiež spolupracuje s inými knižnicami na projekte *Open Library*, ktorý disponuje až 1,6 miliónom kompletných, plne čitateľných kníh.

Archivácia videí

Táto služba poskytuje približne 5 miliónov videí zahŕňajúce celovečerné filmy, rozprávky, krátke filmy a ďalšie. Kategória, ktorá ma najviac zaujala je archívna kolekcia audiovizuálnych nahrávok incidentu z 9/11/2001. Kolekcia obsahuje nahrávky známych vysielacích staníc odo dňa teroristického útoku až po 17. september (viz Obrazok 2).



Obrázek 2: Kolekcia audiovizuálnych nahrávok incidentu z 9/11/2001

Archivácia audio nahrávok
Kolekcia audio nahrávok zahŕňa audio knižky, podcasty, rádiové programy a ďalšie. Počet všetkých archivovaných nahrávok je takmer osem miliónov.

Archivácia softwarových programov
Internet archive disponuje najväčšou zbierkou historických softwarových programov na svete [5]. Zbierka zahŕňa počítačové magazíny, knihy, žurnály, videohry a ďalšie. Kolekcia bola postavená na výnimke zákona o autorských právach, ktorá umožňovala obísť ochranu proti kopírovaniu. Pretože výnimka platí výlučne pre účely archivácie, stiahnutie softwaru nie je možné.

Archivácia obrázkov
Archív obsahuje vyše tri milióny obrázkov rôzneho druhu, od obrazov a malieb, až po obrázky zemského povrchu zo satelitov NASA. Na zbierke sa podieľajú múzeá ako Metropolitné múzeum umenia alebo Brooklynské múzeum.

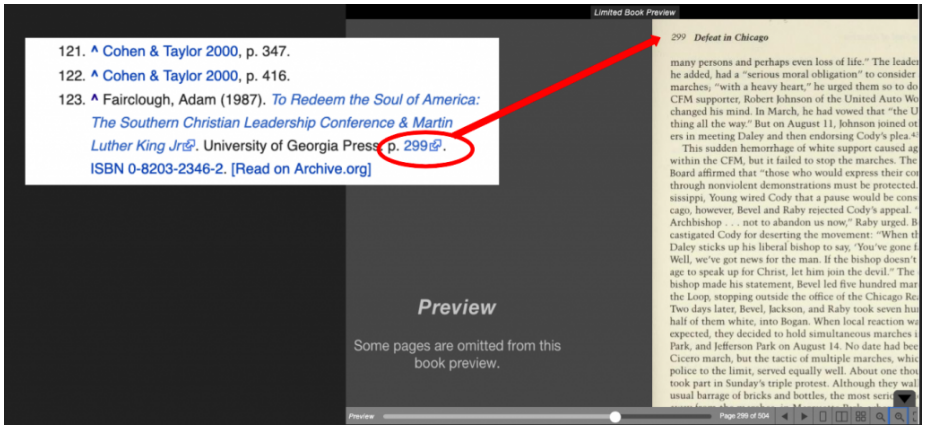
Princíp zaraďovania jednotlivých dokumentov medzi spomínané kategórie je organizovaný prostredníctvom metadát zadaných užívateľom, ktorý daný dokument do systému nahráva, konkrétne atribútom mediatype. Schéma metadát obsahuje mnoho atribútov, ako napríklad jazyk dokumentu, autor, zdroj a mnoho ďalších.

Okrem spomenutých služieb sa organizácia Internet archive zapája do rôznych projektov, ako napríklad archivácia politických televíznych reklám, budovanie skenovacích centier, rozširovanie pokrytia bezdrôtového pripojenia a ďalšie.

5 Najnovšie projekty

Prepojenie wikipédie s Internet archive
Wikipédia predstavuje primárny zdroj prvotných informácií pri budovaní prehľadu o akomkoľvek neznámom tématе, odpovedá na otázky akéhokoľvek druhu a symbolizuje istý druh absolútnej pravdy na internete. Dôvodom, prečo je wikipédia taká rozšírená je skutočnosť, že sa každý fakt opiera o nejakú citáciu, a je teda možné informáciu ľahko dohľadať kliknutím na odkazovanú literatúru. Problém nastáva, ak digitálna podoba literatúry chýba. Napríklad na stránke wikipédie o Martinovi Lutherovi Kingovi Jr. sa nachádzajú citácie ku knihám,

ktoré nie sú dostupné na webe, a je potrebné si zadovážiť ich výtlačok [6]. Iniciatíva organizácie Internet archive sa snaží využiť svoj obrovský archív rôznych kníh a textov na doplnenie chýbajúcej literatúry. Odkazy presmerujú čitateľa na konkrétnu knihu, alebo dokonca na konkrétnu stranu knihy, ktorá odpovedá citácií (viz Obrázok 3). Od počiatku tejto iniciatívy sa podarilo doplniť viac ako stotridsaťtisíc chýbajúcich odkazov [7].



Obrázek 3: Presmerovanie čitateľa priamo k citovanej strane knihy (prevzaté z [7])

Rozšírenie služby Wayback machine
Vylepšenie služby Wayback machine sa týka ukladania webových stránok. Pred spomínaným rozšírením sa ukladala výhradne samotná stránka, no po rozšírení je možné uchovať všetky stránky odkazované danou stránkou. Tento prístup často ukladá namiesto jednej stránky až stovky stránok. Vylepšenie je veľmi užitočné, nakoľko poskytuje väčšiu funkcionalitu a pocit dynamickej stránky [8].

6 Zhodnotenie projektu

Projekt Internet archive je úžasná myšlienka, ktorá umožňuje nazrieť do histórie internetu. Viem si predstaviť, že tento projekt môže byť atraktívny nielen pre historikov a výskumníkov, ale aj novinárov a podobne. Páči sa mi, že sa organizácia zapája do rôznych iných projektov a spoluprácuje s organizáciami ako wikipédia, kongresová knižnica a rôzné ďalšie.

Na druhej strane, knižnica trpí zákonom o autorských právach, ktorý značne limituje dostupnosť aktuálnych kníh a článkov. Avšak cieľom knižnice Internet archive nie je byť konkurencieschopný v komerčnej sfére poskytovania kníh, ale poskytovať archív webu a dokumentov.

7 Metadata v Dublin Core (DC)

```
<dc:title>Internet archive</dc:title>
<dc:creator>Andrej Betík</dc:creator>
```

```
<dc:date>2019-11-20</dc:date>
<dc:description>Esej o aktuálnom stave projektu Internet archive</dc:description>
<dc:type>text</dc:type>
<dc:format>public</dc:format>
<dc:language>sk</dc:language>
```

Reference

[1] About the internet archive. https://archive.org/about/. Accessed: 2019-11-20.

[2] Internet archive history (prelinger archives). https://en.wikipedia.org/wiki/Internet_Archive#History. Accessed: 2019-11-20.

[3] Goals of internet archive. https://www.sciencedirect.com/topics/computer-science/internet-archive. Accessed: 2019-11-20.

[4] Details about internet archive texts and books. https://archive.org/details/texts/. Accessed: 2019-11-21.

[5] Software collection of internet archive. https://en.wikipedia.org/wiki/Internet_Archive#Software/. Accessed: 2019-11-21.

[6] The internet archive is making wikipedia more reliable. https://www.wired.com/story/internet-archive-wikipedia-more-reliable/. Accessed: 2019-11-21.

[7] In the martin luther king, jr. article of wikipedia, page references can now take you directly to the book. https://blog.archive.org/2019/10/29/weaving-books-into-the-web-starting-with-wikipedia/. Accessed: 2019-11-21.

[8] The wayback machine save page now is new and improved. https://blog.archive.org/2019/10/23/the-wayback-machines-save-page-now-is-new-and-improved/. Accessed: 2019-11-21.