# Internet Archive
# http://www.archive.org

Marek Hlaváč, 1. ročník navazujícího magisterského studia

2.12. 2011

# Charakteristika projektu

Knihovny existují, aby zachovaly kulturní artefakty společnosti a zajistily přístup k nim. Bez kulturních artefaktů nemá civilizace žádnou paměť a žádný mechanismus, aby se mohla učit ze svých úspěchů či neúspěchů. Internet Archive se snaží, aby internet – nové médium s velkým historickým významem – a další digitální materiály nezanikly. Komunita Internet Archivu spolupracuje s institucemi včetně Knihovny Kongresu (Library of Congress) či Smithsonian a snaží se uchovat záznamy pro budoucí generace [1].

# Cíle a aktuální stav projektu

Internet Archiv je americká nevýdělečná organizace 501(c)(3), která byla založena za účelem nabízet permanentní přístup výzkumníkům, historikům, vědcům, lidem se zdravotním postižením a široké veřejnosti k historickým sbírkám v digitální podobě. Společnost byla založena v roce 1996 v San Franciscu sponzorována společnostmi jako například Alexa Internet a dalšími. V roce 1999 se organizace začala poměrně rapidně rozrůstat a nyní Internet Archive obsahuje texty, audio, pohyblivé obrázky (animace, filmy, televizní záznamy apod.), software, archivní webové stránky a poskytuje specializované služby pro adaptivní čtení a přístup k informacím pro nevidomé a ostatní osoby se zdravotním postižením. Mimo jiné je Internet Archive členem ALA (American Library Association).

V současné době je velikost webových sbírek Internet Archivu tak velká, že jeho používání vyžaduje znalosti programování, nicméně se vyvíjí nástroje a metody, které umožní široké veřejnosti jednoduchý a smysluplný přístup k naší společné historii. Kromě rozvíjení vlastní kolekce se komunita Internet Archivu snaží podporovat vznik dalších internetových knihoven ve Spojených státech i jinde [1].

# Popis projektu

## Služby

Jaké služby poskytuje Internet Archive? Na webových stránkách archive.org nalezneme v horní části obrazovky horizontální lištu, která ukrývá hlavní nabídku. Po rozkliku každá položka menu odkryje submenu, jenž dále člení obsah dané kategorie dle obsahu či nabízí určité služby.

### *Web*

První záložkou je „Web“ jenž obsahuje nabídky: Wayback Machine, Archive-It, Blog a Heritrix.

- **Wayback Machine** je služba umožňující lidem navštívit archivní verze webových stránek. Návštěvník pouze napíše URL, zvolí datum a může si prohlížet archivní verzi dané webové stránky.
- **Archive-It** je předplatitelná služba od Internet Archivu umožňující vytvářet a uchovávat sbírky s digitálním obsahem. Díky uživatelsky přívětivé webové aplikaci mohou partneři služby Archive-It sklízet, třídit, uspořádávat a procházet jejich archivní kolekce. Kolekce jsou uchovávány v datovém centru Internet Archivu a jsou přístupné prostřednictvím URL a fulltextového vyhledávání [2].
- **Blog:** pod tímto odkazem naleznete jak již říká „anchor text“ blog týkající se archivací webových stránek, na kterém se publikují články již od ledna 2007.
- **Heritrix** je projekt zabývající se open-source rozšiřitelným archivačním webovým robotem (web robot, web crawler). Heritrix je archaické slovo pro dědičku. Protože Heritrix web crawler se snaží sbírat a uchovávat digitální artefakty naší kultury ve prospěch budoucích

generací a výzkumů, zdálo se toto označení příhodné [3].

### *Moving Images*

Toto menu příznačně nazvané „moving images“ obsahuje poměrně mnoho kolekcí rozdělených podle typu a obsahu. Můžeme zde najít následující kolekce: Animation & Cartoons, Arts & Music, Community Video, Computers & Technology, Cultural & Academic Films, Ephemeral Films, Movies, News & Public Affairs, Prelinger Archives, Spirituality & Religion, Sports Videos, Television, Videogame Videos, Vlogs a Youth Media. Tyto kolekce obsahují od stovek až po stacisíce nejrůznějších videoklipů, filmů, dokumentů a podobných „pohyblivých obrázků“ v různých formátech.

### *Texts*

Pod záložkou „Texts“ naleznete následující submenu: American Libraries, Canadian Libraries, Universal Library, Community Texts, Project Gutenberg, Children's Library, Biodiversity Heritage Library a Additional Collections. Již z názvu je u většiny položek patrné co se pod nimi skrývá, tudíž se zmíním pouze o jediné položce a tou je projekt Gutenberg.

**Project Gutenberg** odstartoval v roce 1971, kdy Michael Hart dostal od operátorů přístup k mainframu „Xerox Sigma V“ v laboratoři na univerzitě v Illinois. Tento přístup mu byl umožněn díky účtu, se kterým získal strojový čas v hodnotě 100 000 000 dolarů. Michael prohlásil, že nemůže udělat nic ve smyslu „normálního computingu,“ aby splatil výpočetní čas v takové hodnotě, v jaké mu byl darován a tak musí vytvořit něco takto hodnotného jiným způsobem. O hodinu a 47 minut později prohlásil, že největší hodnota vytvářená počítači by neměl být jejich výpočetní čas, ale získávání, ukládání a vyhledávání toho, co bylo uloženo v našich knihovnách. Po zdigitalizování Deklarace nezávislosti spojených států řekl, že právě „vydělal“ 100 000 000 dolarů, jelikož tato kopie může být v budoucnosti součástí elektronických knihoven 100 000 000 počítačových uživatelů. Tímto odstartoval projekt Gutenberg nazvaný podle vynálezce knihtisku. V současné době nabízí tento projekt zdarma ke stažení přiližně 36 000 elektronických knih v několika formátech (například ePub, mobi – Kindle, HTML, prostý text a další) [4].

### *Audio*

Audio knihovna Internet Archivu obsahuje více než 200 tisíc volných digitálních nahrávek nejrůznějšího původu, jak můžete posoudit dle submenu, které obsahuje následující položky: Audio Books & Poetry, Community Audio, Computers & Technology-Audio, Grateful Dead, Live Music Archive, Music & Arts, Netlabels, News & Public Affairs, Non-English Audio, Podcasts, Radio Programs, Spirituality & Religion. Mnohé z těchto nahrávek jsou dostupné zdarma ke stažení.

### *Software*

Softwarový archiv je určen k uchování a k umožnění přístupu ke všem druhům vzácných, těžko nalezitelných, legálně stažitelných softwarových titulů a k základním informacím o těchto titulech.

Kolekce zahrnuje širokou škálu materiálů souvisejících se softwarem včetně sharewaru, freewaru, videí prezentujících softwarové tituly, promo videí a preview počítačových her, tabulek s „high-score“ a „replays“ nejrůznějších herních žánrů a umění filmové tvorby za pomoci enginů počítačových her. Tato nabídka obsahuje následující submenu: DigiBarn, The Shareware CD Archive, Tucows Software Library, The Vectrex Collection.

## Projects

Projekty na kterých se podílí či je přímo vytvořila komunita Internet Archivu:

- **Open Library** – Jedna webová stránka pro všechny vydané knihy. Open Library je otevřený projekt na kterém se může kdokoliv podílet. Software je volně dostupný, stejně tak jako

webové stránky, data i dokumentace.

- **Wayback Machine** – viz Služby – Web
- **Archive-It** – viz Služby – Web
- **Scanning services** – Internet Archiv může digitalizovat Vaše sbírky, permanentně je uchovat a umožnit k nim volný přístup. Internet Archiv již naskenoval více než 200 milionů stránek s partnery od „Library of Congress“, přes „Smithsonian“ k „New York Public Library“, Harvardu a MIT.
- **BookServer** – Projekt BookServer poskytuje otevřenou architekturu pro prodej, půjčování a distribuci knih přes internet. Postavený na otevřených standardech, umožňuje BookServer model široké síti nakladatelství, knihkupectvím, knihovnám a dalším stranám zpřístupnit katalogy knih čtenářům přímo prostřednictvím jejich notebooků, telefonů, netbooků či elektronických čteček.
- **NASA Images** – Tento projekt vznikl ve spolupráci Internet Archivu a NASA. Má přinést veřejnosti přístup ke kolekcím fotografií, videí a audio záznamů NASA. NASAImages.org je největší kolekce mediálních dat z NASA, dostupná a prohledávatelná z jedné jediné stránky. Tato webová stránka obsahuje vše od klasických fotografií až po vzdělávací programy či HD videa a stále se přidávají ať už nová či archivní média ze všech center NASA. Komunita tohoto projektu doufá, že prostřednictvím rozvoje nasaimages.org podpoří vzdělávání, usnadní stipendia matematických a přírodních věd na všech úrovních a také vytvoří obecný zájem o vesmír, letectví a astronomii.
- **Open Content Alliance (OCA)** – OCA je výsledkem společného úsilí skupiny kulturních, technologických, neziskových a vládních organizací po celém světě. Tato aliance pomáhá vytvářet archivy s vícejazyčnými digitalizovanými texty a multimediálními materiály. Archiv s těmito materiály je přístupný na stránkách Internet Archivu, přes Yahoo! a další vyhledávače či webové stránky.
- **Education** – „Open Education Resources library“ obsahuje stovky bezplatných kurzů, videopřednášek a doplňujících materiálů z univerzit v USA a v Číně.
- **Bookmobile** – Bookmobile je mobilní digitální knihovna schopná stahovat veřejně dostupné knihy z internetu prostřednictvím satelitu a tisknout je kdykoliv, kdekoliv a pro kohokoliv. Bookmobile cestuje napříč Spojenými státy a podobná auta byla sestrojena a využívána také v Egyptě a Ugandě.
- **Open Community Networks** – Tento projekt vznikl v roce 1997 a od té doby se velmi rozrostl. Poskytuje bezplatný vysokorychlostní drátový a bezdrátový internet obyvatelům San Francisca s nízkým příjmem.
- **Petabox** – Petabox byl speciálně navržen pracovníky Internet Archivu pro účely bezpečného ukládání a zpracování jednoho petabajtu (1 milion GB) informací. Cílem bylo vytvořit levný úložný systém s nízkým výkonem, vysokou hustotou, snadno škálovatelný a udržovatelný. Petaboxy jsou nyní v provozu v hlavních akademických institucích a vládních agenturách. Budovy Internet Archivu mají nyní úložiště velikosti 10 petabajtů zprostředkované PetaBoxy, jejichž počet se stále zvětšuje.
- **301Works.org** – 301works.org je nezávislá služba pro archivaci URL mapování. Cílem služby je poskytnout ochranu pro každodenní uživatele krátkých URL služeb zajištěním transparentnosti a trvalosti jejich mapování.

### Úložiště

Uložení archivu sbírky zahrnuje analýzu, indexování a fyzické kódování dat. Vzhledem k tomu, že internetové sbírky rostou exponenciálním tempem, tento úkol představuje trvalou výzvu.

Hardware Internet Archivu se skládá z PC s clustery s IDE disky. Data jsou uložena na DLT pásky a pevné disky v různých formátech v závislosti na kolekci. Webová data jsou přijata a uložena v archivním formátu 100-magabajtových ARC souborů poskládaných z jednotlivých souborů. Alexa Internet navrhuje ARC jako standard pro archivaci internetových objektů [1].

### Uchovávání (Konzervace)

Úkolem konzervace je trvalým způsobem chrátnit uložené prostředky před poškozením či zničením. Hlavním problémem je ochrana před následky nehod, degradace dat, zachování dostupnosti dat a zastarávání formátů.

**Nehody:** Jakákoliv média či úložiště pro ukládání dat jsou potencionálně ohrožena havárií nebo přírodní katastrofou. Vytvoření kopií kolekcí Internet Archivu na více místech může toto riziko zmírnit. Součást kolekce je již takto ošetřena a komunita Internet Archivu pracuje tak rychle jak je to jen možné, aby udělala to samé se zbytkem kolekcí.

**Migrace:** Postupem času mohou paměťová média degradovat až k bodu, kdy již nebude možné data na těchto médiích přečíst. I když DLT páska je dimenzována na 30 let, průmyslovým pravidlem je migrace dat každých 10 let. Internet Archive již nevyužívá pásky, místo toho používá takzvaný Petabox, systém, o němž se zmiňuji v souvislosti s projekty Internet Archivu [1].

## Vlastní zhodnocení projektu

Internet Archive je bezpochyby jeden z nejzajímavějších projektů svého druhu a má velký potenciál k dalšímu rozvoji. Komunita Internet Archivu je velmi iniciativní a spolupracuje s poměrně velkým množstvím organizací na nejrůznějších projektech. Fakt, že se tato komunita přímo nesoustředí pouze na digitalizaci knih, ale snaží se digitalizovat a uchovat informace nejrůznějšího typu, dělá Internet Archiv portálem disponujícím ohromným množstvím informací, kde můžeme vyhledávat data, která si pak většinou zdarma stáhneme ve formě, která nám právě vyhovuje. Aktivity Internet Archivu se spíše rozrůstají, než aby uvadaly, a tak se těším, s jakými dalšími zajímavými projekty tato komunita v budoucnosti přijde.

## Literatura


1. About the Internet Archive [online], [cit. 21.11.2011]. Dostupné z: <http://www.archive.org/about/about.php>
2. Internet Archive Projects [online], [cit. 21.11.2011]. Dostupné z: <http://www.archive.org/projects/>
3. Heritrix [online], [cit. 22.11.2011]. Dostupné z: <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>
4. The History and Philosophy of Project Gutenberg by Michael Hart [online], [cit. 22.11.2011]. Dostupné z: <http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart>


## Metadata v DC

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" /><meta name="DC.Title" content="Open Library" />

<meta name="DC.Creator" content="Marek Hlaváč" />

<meta name="DC.Language" content="cs-CS" />

<meta name="DC.Date" content="2011-12-02" />

<meta name="DC.Description" content="Informace o projektu Internet Archive" />

<meta name="Type" content="Text" />

<meta name="DC.Format" content="application/pdf" />

<meta name="DC.Source" content="http://www.archive.org/about/about.php" />

<meta name="DC.Source" content="http://www.archive.org/projects/" />

<meta name="DC.Source" content="https://webarchive.jira.com/wiki/display/Heritrix/Heritrix" />

<meta name="DC.Source" content="http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart" />