Petr Švenda
xsvenda@fi.muni.cz
6. ročník, 12/2003

Internet Archive

http://www.archive.org/

Internet Archive je americká nezisková organizace založená roku 1996 s cílem vytvořit digitální knihovnu volně přístupnou výzkumným a vzdělávacím účelům, obsahující archiv digitálních dokumentů sítě Internet, textových kolekcí, volně přístupných hudebních a filmových nahrávek a další materiály existující pouze v digitální podobě („born-digital“). Hlavní část archivu tvoří dokumenty sbírané společností Alexa Internet od roku 1996 z Internetu. Druhou část tvoří data získána formou darů od různých organizací a jednotlivých dárců.

Užití Internet Archivu

Myšlenka archivace Internetu je založena na poznání, že digitální data presentovaná pomocí sítě Internet, patří k naší kulturní historii, stejně jako dokumenty tištěné. Proto je vhodné je uchovávat nejen pro naši aktuální potřebu, ale i pro potřebu budoucích pokolení. Je tedy snaha uchovávat co nejvíce informací, bez ohledu na aktuální praktické využití a zachovat tak pro budoucí generace co nejvěrnější obraz současnosti.
Jde tedy především o tyto cíle:

- *Sledování vývoje internetu* – periodická záloha stavu Internetu poskytuje podklady pro sledování a statistické vyhodnocování trendů změn obsahu, například rozšíření konkrétních ekonomických modelů, rozšíření a rozsah reakcí na konkrétní problémy (Y2K, ekologické uvědomění, války…) nebo vývoje popularity jednotlivých vědních oborů.
- *Archivace politické historie* – pro potřeby studia fungování a vývoje politické kultury je třeba zajistit přístup k dokumentům a prohlášením, které byly aktuálně zveřejňovány. Pokud jsou tyto informace přístupné pouze v digitální podobě, je třeba je archivovat, neboť klasické knihovny již neposkytují možnost zpětně tyto informace získávat. Nedemokratické státy často vykazují snahu o omezení a kontrolu přístupu k Internetu a cenzuru umisťovaného obsahu. Takové jednání může být v archivu z dlouhodobé perspektivy patrné.
- *Sledování jazykových změn* – archivované digitální dokumenty umožňují sledovat změny a vývoj gramatiky, slovní zásoby konkrétního jazyka a jeho rozšíření, tak jako v minulosti tištěné knihy a časopisy.
- *Nalezení mrtvých odkazů* – vytvořený archiv je možno použít pro získání stránek, které již nejsou v současné době na původní adrese dostupné nebo neobsahují původní informace. Originální verze stránek lze v archivu i odkazovat a zajistit tak větší trvanlivost a validitu odkazů.

Principy sběru stránek

Archivované stránky jsou tvořeny především daty poskytnutými organizací Alexa Internet [2], která se zabývá sběrem internetových stránek od roku 1996 pro účely odhadu vývojových trendů internetu pro komerční využití serverem Amazon. Alexa provádí pravidelný sběr pomocí softwarových robotů, kteří systematicky procházejí stránky na základě odkazů z dříve navštívených stránek. Ze způsobu procházení vyplývá, že nedochází k archivaci stránek, které nejsou volně přístupné z důvodu nutnosti autentizace, nejsou referovány žádnou další stránkou (Orphan pages) nebo o nich Alexa „neví". Archivaci takové stránky lze v některých případech zajistit přímou žádostí o sběr na stránkách Alexy. Alexa robot navíc respektuje direktivy uvedené v robots.txt (Standard for Robot Exclusion [5]) a neprovádí sběr stránek, u kterých je to příslušnou direktivou zakázáno. Není prováděn sběr e-mailové korespondence ani chatu, tedy privátních informací uživatelů.

Nejlépe je prováděn sběr a archivace statických HTML stránek, které neobsahují skripty a formuláře. V případě dynamicky generovaných stránek probíhá sběr snadno, pokud je výsledný formát opět čisté HTML. Pokud stránka obsahuje skripty nebo formuláře vyžadující komunikaci s mateřským serverem, není tato funkcionalita v archivu obsažena. Problémy při archivaci také způsobují skripty, které doplňují pouze relativní adresu pro odkazy umístěné na stránce, nikoli celou URL adresu.

Way-Back Machine (WBM)

Data získaná sběrem jsou po zhruba šesti měsících zpřístupněna prostřednictvím rozhranní WBM[7], které umožňuje zadat adresu požadované stránky a případně i čas výskytu původní stránky. Pokud je stránka v archivu obsažena, je zobrazena verze stránky archivována v tomto rozmezí. Pokud není období specifikováno, zobrazí se seznam všech archivovaných verzí hledané stránky dle data archivace. Odkaz do archivu na výslednou stránku může být použit pro běžné odkazovaní v případě, že původní stránka již není na původní adrese dostupná.

WBM obsahuje v současnosti zhruba 300 TB dat na několika desítkách serverů, měsíční nárůst je v současné době asi 12 TB dat. Celý archiv IA je provozován na linuxových x86 serverech s 512 MB RAM a 1TB ATA diskového prostoru.

Pro vyhledávání a práci s archivem je použito třídimenzionálního indexování (technologie od Alexa Internet), která indexuje jednotlivé záznamy i v čase.

Vyhledávání v Internet Archivu

Pokud nepostačuje prohledávání archivu pomoci WBM, lze použít službu Recall [8], která je čerstvě k dispozici v beta verzi a nabízí inteligentní fulltextové vyhledávání v 11 miliardách archivovaných stránkách. Indexování bylo provedeno pomocí technologie CobWebSearch [9] vyvinuté na Standfordské universitě s výslednou indexovací tabulkou o velikosti 2 terabajty. Tato technologie poskytuje možnost automatického určení kategorie a tématu pro indexovanou stránku (v současné době přes 50 tisíc kategorií), řazení (ranking) výsledků dotazu na základě obsahu stránky, nikoli na základě její popularity. Zajímavou možností je kontinuální personalizace řazení na základě oblastí zájmů uživatele, kdy je snaha na základě předchozích dotazů konkrétního uživatele určit obor jeho zájmu a stránky korespondující s určeným oborem pak řadit ve výsledcích dříve.

Další možností je užití programového rozhraní archivu, v současné době (12/2003) však z důvodu úprav není možné provést novou registraci.

Ukládání, archivace a vyhledávání

Vzhledem k odlišné povaze digitální informace je zároveň nutno použít i odlišné metody archivace. Především je nutno řešit problémy vyplývající z omezené životnosti médií určených pro zaznamenávání digitálních dat, zastarávání čtecích zařízení a vývoj formátů použitých pro jejich representaci.
Archivovaná data jsou ukládána na IDE disky a DLT pásky ve formátu ARC [4] v souborech o velikosti 100 MB. Životnost DLT pásek je zhruba 30 let, z důvodů zastarávání čtecích zařízení se však migrace na nové médium plánuje v kratší době. Pro ochranu archivovaných dat před poškozením je kopie části archivu umístěna na geograficky odlišném místě a v dohledné době by měl být takto zálohován celý archiv. Formát ARC je navrhován jako standard pro dlouhodobou archivaci internetových objektů. Zároveň budou v archivu uchovávány i emulátory prostředí a software pro přístup k již nepoužívaných a zastaralým formátům.

Jak zajistit archivaci stránky

Pokud není stránka obsažena ve WBM, lze její sběr zajistit přímou žádostí u Alexy. Jednou z možností je instalace Alexa Toolbar [3], který automaticky zajistí archivaci každé navštívené stránky, zároveň však uchovává poměrně velké množství privátních informací (obsahy vyplňovaných formulářů apod.). Další možností je přímé zadání adresy stránky, u které je požadována archivace, do formuláře na stránkách Alexy [10]. Stránka bude archivována nejpozději do 48 hodin a do WBM bude umístěna za 6 měsíců.

Odstranění stránky z archivu

V případě, že majitel stránky nemá zájem o uchování své stránky v archivu IA, může umístit soubor robots.txt do kořenového adresáře domény s direktivami:
User-agent: ia_archiver
Disallow: /
Tato operace způsobí, že nebude prováděn nový sběr stránek z příslušné domény a zároveň dojde k retrospektivnímu odstranění stránek z archivu. Spolu s možností retroaktivního odstranění stránek z archivu však vzniká problém související s tzv. CyberSquatem, při kterém dojde k možnosti modifikace stránek novým majitelem, který nemá žádný vztah k původnímu tvůrci stránek, například po zániku firmy a uvolnění příslušné domény. Nový majitel pak žádostí o odebrání stránek z archivu způsobí nedostupnost i dříve archivovaných stránek, které mu nepřísluší.

Archivace hudebních a filmových nahrávek (Prelinger, Live Music Archive)

Kromě archivace internetových dokumentů poskytuje IA možnost archivace živých audio nahrávek v sekci zvané „Live Music Archive“ [11]. Uchovávané nahrávky jsou volně stažitelné v neztrátových kompresních formátech. Archiv živých nahrávek je sestaven převážně z darů hudebníků, kteří si přejí uchování svých veřejných projekcí.
Archiv obsahuje a zpřístupňuje filmové nahrávky z Prelingerova archivu [6], ve kterém je obsaženo přes 4000 převážně amerických filmových nahrávek z vzniklých před rokem 1964. Na základě zákona o ochraně autorských práv, platného od roku 1964, jsou

všechny nově vzniklé nahrávky automaticky chráněny a není je možné zpřístupnit bez souhlasu autora. Nahrávky jsou uchovány ve formátu MPEG-2 v rozlišení 480x480 nebo 368x480 pixelů.

Archivace textových archivů (Project Gutenberg, Million Book Project, Arpanet …)

Projekt Gutenberg byl započat v roce 1971 na universitě Illinois a je postaven na myšlence, že největším přínosem počítačů není jejich výpočetní kapacita, ale schopnost archivace a efektivního vyhledávání. Cílem projektu je převést klasická, volně přístupná knižní díla do elektronické podoby s použitím co nejlépe přenositelného formátu, za který byl zvoleno klasické ASCII kódování. V současné době má archiv takřka 10 tisíc volně dostupných titulů.
Cílem projektu „Million Books Project“ je převedení jednoho milionu knih elektronické podoby do roku 2005 a vytvoření veřejně přístupné elektronické knihovny s příslušnou funkcionalitou. Pilotní projekt obsahující přes 10 tisíc knih je již také volně přístupný.
Textová kolekce Arpanet obsahuje memoranda, rozhovory, periodika a ostatní materiály dokumentující vývoj „Advanced Research Projects Agency Network“, předchůdce internetu.
Textová kolekce „Dance Archiv“ obsahuje digitalizovanou podobu několika stovek učebnic tance, publikovaných v letech 1415 až 1920.

Problém copyrightu na archivované informace

Archivují se pouze veřejně přístupné informace, není prováděna archivace e-mailové komunikace, chatu, stránek chráněných heslem a jsou respektovány direktivy uvedené, v souboru robots.txt. Předpokládá se tedy implicitní souhlas autorů s uchováním jejich zveřejněných dokumentů a v případě jejich žádosti jsou informace retroaktivně odstraněny z archivu. Pro možnost archivaci počítačových programů jsou důležité výjimky z DMCA (Digital Millenium Copyright Act), udělené v říjnu 2003, které umožňují legálně pořizovat záložní kopie obcházející ochranné prvky programů v případě, že se jedná o již nepoužívaný nebo nefunkční formát ochrany a samotný program je již zastaralý.

Závěr a hodnocení

Internet Archivu vnímám jako velmi přínosnou iniciativu a to i když pominu možné přínosy pro budoucí generace historiků, které nejsem v současné době plně ohodnotit. Především se jedná o praktickou možnost hledání již nedostupných dokumentů, kterou jsem v krátké době několikrát s úspěchem použil. Vhodným doplněním současných hypertextových odkazů by mohl být i odkaz na cílový zdroj do IA, což je ale vhledem k prodlevě v zpřístupňování archivovaných dokumentů v současnosti prakticky nepoužitelné. Tato prodleva je jedním z mála záporů, na které jsem narazil. Zajímavou částí projektu, a vzhledem k množství záznamů snad i nutnou, je způsob prohledávání archivovaných záznamů službou Recall, která se snaží zlepšit efektivitu klasického fulltextového prohledávání strojovým porozuměním obsahu prohledávaných dokumentů a zadaného dotazu. Pro širší rozšíření je dobrým předpokladem bezplatný přístup k poskytovaným službám.

Literatura

[1] Internet Archive, http://www.archive.org
[2] Alexa Internet, http://www.alexa.com
[3] Alexa Toolbar, http://download.alexa.com/
[4] ARC Format Specification, http://pages.alexa.com/support/arcformat.html
[5] Standard for Robot Exclusion, http://www.robotstxt.org/wc/norobots.html
[6] Prelinger Archive, http://www.prelinger.com
[7] Internet Archive Way-Back Machine, http://web.archive.org/
[8] Internet Archive Recall, http://recall.archive.org/
Personalizovaná verze, http://myrecall.archive.org
[9] CopWebSearch Presentation, http://ia00406.archive.org/cobwebsearch.ppt
[10] Alexa Crawl Page, http://pages.alexa.com/help/webmasters/index.html#crawl_site
[11] Live Music Archive, ftp://etree0X.archive.org.
[12] Million Books Project,
http://www.archive.org/texts/collection.php?collection=millionbooks