Luděk Novotný
Bc. Ap. Inf, ročník 4
12. prosince 2004

Internet Archive - celosvětový archiv webu

V dnešní době stále více kulturních artefaktů vzniká pouze v elektronické podobě (born digital). Projekt Internet Archive si klade za cíl zachovat toto elektronické dědictví pro příští generace.

Úvod

Internet jako informační médium příliš nepodléhá cenzuře a odráží tak skutečnou situaci ve společnosti. Projekt Internet Archive (dále jen archiv) si klade za cíl zachovat co největší část obsahu webu pro budoucí generace. Periodicky jsou ukládány celé obsahy stránek povrchového webu, a to bez ručního výběru, které stránky se budou archivovat a které ne. Díky tomu archiv poskytuje cenné informace pro vědce, historiky a studenty, kteří chtějí porozumět historii.

Archiv založil roku 1996 Brewster Kahle (autor vyhledávače Alexa Internet). V současné době je spravován neziskovou organizací Internet Archive. Finance na provoz a další rozvoj získává formou darů. Mezi nejvýznamnější patří Alexa Internet, AT&T Research, HP Compaq, Kahle/Austin Foundation, Prelinger Archives, Quantum DLT a Xerox PARC. Spolupracuje též s Library of Congress (dále jen LoC), National Science Foundation a společností Smithsonian.

Tradiční velké knihovny (například LoC) mají tisíce zaměstnanců. Naproti tomu počet lidí zajišťujících provoz archivu se pohybuje okolo padesáti. To je mimo jiné možné díky vyšší míře automatizace (automatické generování metadat oproti náročnější ruční tvorbě).

Velikost archivu

Archiv začal vznikat od roku 1996. Pro první rok provozu je charakteristický omezený počet archivovaných stránek a malý počet kopií. Původní frekvence vytváření snímků aktuálního stavu Internetu byla několikrát ročně. Dnes je doba mezi jednotlivými cykly okolo dvou měsíců a řada webů se archivuje týdně až denně.

Průměrná doba po kterou je stránka na Internetu se pohybuje mezi 80 až 100 dny. Toto relativně malé číslo je ovlivněno jednak zanikajícími prezentacemi na úrovni domén druhého řádu a jednak zrušením stránek hostovaných u poskytovatelů prostoru pro webové prezentace zdarma, případně zánikem těchto poskytovatelů.

Poskytovatelé se snaží odstranit stránky u kterých po určitou dobu nedojde k aktualizaci jejich obsahu. Vedlejším efektem tohoto přístupu k uvolňování místa na serverech je nevratný zánik mnohdy jedinečných zdrojů informací.

Příkladem může být příběh ze života, kdy mi web shrnující českou scénu okolo dříve rozšířeného osmibitového počítače ZX Spectrum doslova zmizel před očima. Během jeho procházení mi server, na kterém byl uložen, náhle oznámil jeho nedostupnost z důvodu odstranění.

Přestože stránky již nejsou udržovány, mohou obsahovat informace, které se již nikde jinde na Internetu nevyskytují. Tímto se nenávratně ztrácejí jedinečné informace svázané s určitým, z dnešního pohledu nezajímavým, tématem, které mohou být z historického hlediska cenné.

V roce 2001 archiv indexoval přes deset miliard stránek, které by jinak byly ztraceny.
(Hamelinck, 2001)

Další problém představuje duplicita. Jen u 50 % stránek je provedena změna od doby jejich poslední návštěvy před několika měsíci. To podle (Archive, 2004) znamená, že velká část obsahu uchovávaného v archivu je duplicitní.

Přestože archiv je schopen detekovat výskyt změny stránky[1], při každé návštěvě ji opět kompletně uloží. Důvodem mohou být, kromě nedostupnosti některých jejích částí při pořizování předešlé kopie, především neustálé změny v programu na sklízení webu. Kromě opravování nalezených chyb, jsou jeho změny vyvolány jako reakce na stále se měnící technologie na poli webových prezentací.

Následující údaje ilustrují rostoucí velikost archivu v čase:[2]

- archiv obsahuje minimum z roku 1996 a nic před ním (ani před Internetem: bbs, gopher),
- 2001: velikost archivu 100 TB, rychlost růstu 10 TB za měsíc, 10 miliard stránek,
- 2003: velikost archivu 100 TB, rychlost růstu 12 TB za měsíc,
- 2004: velikost archivu 1 PB, rychlost růstu 20 TB za měsíc, 30 miliard stránek.

Původně bylo 100 TB dat uloženo na 200 paralelně zapojených počítačích (modifikované x86 servery s ATA disky; každý s kapacitou 1 TB). Později archiv vyvinul vlastní zálohovací zařízení „Petabox“ k uchovávání velkého množství dat s minimálními nároky na prostor a s malou spotřebou.

Takovéto množství digitálních dat představuje dle (Hamelinck, 2001) největší digitální knihovnu na světě. Větší než všechny knihovny, včetně LoC, dohromady.

Archiv oproti vyhledávačům (Google, Yahoo!) ukládá kompletní stránky včetně obrázků (naprostá většina ve formátu GIF, JPEG). Ukládány jsou též dokumenty (PDF, DOC), audio, video a binární soubory. Problém představují stránky, které vyžadují ke správnému zobrazení komunikaci se serverem, používají jazyk JavaScript nebo Java applety.

Pro archivaci metadat archiv používá vlastní formát (ARC) založený na XML. V rámci něho jsou použita též metadata ve formátu Dublin Core.

Ochrana takového množství dat před technickým selháním nebo přírodní katastrofou představuje netriviální problém. Současná strategie spočívá ve vytvoření několika identických kopií archivu po celém světě. První je umístněna v San Franciscu ve Spojených státech, druhá vznikla v Egyptě v Alexandrii (Bibliotheca Alexandrina; vznikla na počest slavné Alexandrijské knihovny). Jsou plánována další dvě sídla – první ve střední Evropě a druhé v Asii.

Kolekce

Pro klasické knihovny je typické, že své sbírky uspořádávají do kompaktních kolekcí, které jsou následně zpřístupňovány. Archiv ve spolupráci s jinými subjekty začíná vytvářet kolekce na vybrané historické události (Y2K, 11. září).

Kromě vlastních snímků stavu webu a sítě Usenet archiv hostuje řadu dalších projektů. V rámci nich jsou přístupné kolekce filmů z let 1903–1973, záznamy televizního vysílání, hudby (především záznamy koncertů), textů (například Open Source Books, projekt Gutenberg) nebo archiv software.

[1] Ve výsledku vyhledávání signalizováno hvězdičkou.

[2] Zdroj: (Hamelinck, 2001), (Bibliotheca Alexandrina), (Kahle, 2002).

Zpřístupnění

Ačkoli je archiv budován od roku 1996, jeho obsah byl zpřístupněn až v říjnu 2001. Vlastní přístup je realizován prostřednictvím webového rozhraní „Wayback Machine“ – vyhledávací technologie věnovaná archivu společností Alexa Internet.

Wayback Machine na základě URL (*Uniform Resource Locator*) a volitelně časového intervalu vrací seznam dostupných archivovaných kopií stránky. Archiv uvádí, že v roce 2002 zpracovával průměrně přes 200 dotazů za sekundu. (Kahle, 2002)

Vyhledávání na základě klíčových slov, tj. bez nutnosti znalosti přesné adresy, umožňovala experimentální služba „Recall Search“. Byla spuštěna v září 2003 a o rok později zastavena. V budoucnu by se měla opět objevit.

Jestliže není k dispozici žádná archivní kopie stránky (a nejedná-li se o úmyslné omezení ze strany poskytovatele obsahu), bude sklizena do 24 hodin a poté při každém dalším sklízení webu.

Data jsou v archivu uložena a následně zpřístupněna v takovém stavu, v jakém byla získána z původního umístnění na Interneru. Díky tomu, že dnešní webové prohlížeče mají dostatečnou podporu pro zobrazování starých stránek, není potřeba provádět úpravu dat nebo jejich migraci.

Vyskytují-li se v dokumentech odkazy s absolutní adresou, tak jsou zachovány. Při vlastním prohlížení jsou takovéto adresy odkazů a obrázků automaticky přesměrovány na časově nejbližší verzi uloženou v archivu.[3] Poznamenejme, že odkazy u stránek uchovávaných vyhledávači (Google, Yahoo!) směřují na své původní umístnění a neumožňují tedy procházení na straně archivu vyhledávače.

Robot

Archiv, stejně jako dnešní vyhledávače, používá ke sklízení povrchového webu roboty (též crawler, spider nebo agent). Řada poskytovatelů obsahu si nepřeje, aby jejich stránky byly automaticky zpracovávány. Toho lze dosáhnout za pomoci SRE (*Standard for Robot Exclusion*). Stačí do adresáře serveru umístnit soubor `robot.txt` s následujícím obsahem:

```
User-agent: *
Disallow: /
```

a „slušné“ nástroje na hromadné ukládání webu se tímto přáním autorů stránek budou řídit. Archiv takto označené stránky dále neukládá a stávající snímky znepřístupní.

Brewster Kahle (Kahle, 2002) říká, že takto blokovaných stránek je minimum. Problém představuje spíše typ těchto zdrojů. Zpravidla se jedná o vydavatele novin, fotografy případně osobní stránky.

Za zmínku stojí, že samotný archiv blokuje případné automatické zpracování svého obsahu. (Lze požádat o výjimku.)

Bibliografické citace

Neexistuje jednotný způsob citací. Na území naší republiky platí norma ČSN ISO 690, ale řada odvětví preferuje své vlastní zvyky, jak uvádět bibliografické citace. Pomineme-li různé pořadí jednotlivých prvků, má většina způsobů společné to, že u elektronicky dostupných zdrojů uvádí jak adresu (URL), tak datum, kdy byl daný zdroj na tomto místě k dispozici.

K přirozenému požadavku u citací patří možnost identifikace původního citovanému materiálu. U tradičních zdrojů je možnost použít knihovny. Z praktických důvodů ve světě roste požadavek po perzistentní identifikaci (a dostupnosti) digitálního zdroje.

[3] Pro správnou funkci odkazů a zobrazování obrázků je nutné mít v prohlížeči zapnutou podporu pro JavaScript.

Realizace perzistentního identifikátoru jako uspořádané dvojice (adresa, datum) má oproti nepřímé adresaci (například handle nebo DOI) řadu výhod:

- K identifikaci není potřeba, aby zdroj byl zaregistrován u příslušné autority (rezoluční služby). Většina zdrojů nemá přidělen takovýto identifikátor.
- Na základě identifikátoru nezískáme „pouze“ metadata ale obvykle vlastní dokument.[4]
- Lze vytvářet perzistentní odkazy přímo do archivu. (Určitá garance dostupnosti dokumentu i v případě jeho zániku na původním umístnění.)
- Číselné identifikátory se špatně pamatují (jsou ovšem vhodné k automatickému zpracování).
- Identifikace není podmíněna existencí právě jedné rezoluční autority. V případě zániku organizace (například neprodloužením grantu) spravující nepřímé identifikátory je obtížná jejich další rezoluce. V případě replikace můžeme identifikátor (adresa, datum) rezolvovat prostřednictvím libovolného archivu.
- Není potřeba zavádět další identifikátor. Jak adresa, tak datum jsou dnes běžně uváděny jako součást bibliografických citací.

Jestli dva dokumenty stejného obsahu, ale s jinou adresou, jsou identické, je otázkou konvence (a netriviální složitosti implementace rezoluční služby na straně archivu).

Právní aspekty

Většina filmů, knih a zvukových záznamů obsažených v archivu spadá do kategorie public domain.

Prakticky všechny kopie textů a obrázků archiv vytváří bez písemného svolení. Pokud nechcete, aby archiv vytvářel kopie vašeho obsahu, můžete být vyjmuti ze seznamu použitím SRE (stávající kopie ovšem zůstanou v archivu; budou jen nepřístupné) nebo kontaktovat právníky archivu a požadovat odstranění stávajících kopií.[5]

„Kromě toho není zřejmé, zda je legální, aby crawler kopíroval web bez svolení; Alexa Internet aktivně kopíruje, ale webové stránky odstraňuje z archivu na žádost tvůrce nebo držitele autorských práv (strategie předpokládaného souhlasu).“

(Lyman, 2003)

Archiv je nezisková společnost ale poznamenejme, že řada vyhledávacích služeb (Google, Yahoo!) nevyhledává v Internetu, ale ve svých kopiích webu (obsahují autorsky chráněné texty a obrázky), které budují pro komerční účely.

Archiv se snaží chránit zájmy držitelů autorských práv několika kroky:

- zpřístupněny jsou jen stránky starší než šest měsíců,
- vyřazení na žádost (nepořizovat další kopie, případně znepřístupnit nebo smazat kopie stávající),
- možnost omezení přístupu k archivu.

Jeden z největších problémů archivů představuje neexistence obdoby povinného výtisku u knih pro elektronické dokumenty. Je způsobena nejednotností autorské ochrany v jednotlivých zemích a chybějícím řešením na mezinárodní úrovni. Díky tomu právní otázka brzdí další rozvoj technologií.

[4] Ručně vytvářená metadata jsou zpravidla kvalitnější, než automaticky generovaná na straně archivu (v případě nedostupnosti archivní kopie například z důvodu ochrany autorských práv).

[5] Požadují ovšem vyplněnou a podepsanou žádost, ve které se zavazujete, že stránka je opravdu vaše a nechcete, aby byla dále zpracovávána. Na druhou stranu jakýkoli materiál uložený v archivu nesmí být kopírován bez výslovného povolení jeho provozovatele.

Závěr

Nepodařilo se mi nalézt obdobný projekt, který by měl za cíl archivovat kompletní obsah webu. Existuje ovšem řada národních projektů: Pandora (Austrálie), European Web Archive, Xyleme (Francie), Nordic Web Archive nebo WebArchiv (Česká republika). Řada archivů je též zaměřena na omezenou tématiku (například archivace webu státních institucí).

Archiv poskytuje čistě praktický nástroj k rezoluci bibliografických citací z digitálních dokumentů. Umožňuje nám i budoucím generacím pohlédnout proti proudu času zpět do historie a tím zviditelnit skryté souvislosti.

Možná největší přínos archivu představuje nutnost hledat odpovědi na nové otázky – jak spolehlivě uchovávat a zpřístupňovat „nepředstavitelné“ objemy dat – a z toho plynoucí výzkum a vývoj nových technologií, které mohou najít své uplatnění nejen v digitálních archivech, ale v celé řadě jiných oblastí lidské činnosti.

Zdroje

Archive, *Internet Archive*, [on-line]. 2004. [cit. 2004-12-02] Dostupné na Internetu: (http://www.archive.org).

Bibliotheca Alexandrina, textitInternet Archive, [on-line] [cit. 2004-12-02] Dostupné na Internetu: (http://www.bibalex.org/english/initiatives/internetarchive/about.htm).

Hamelinck A., *Internet Archive Launches Wayback Machine*, JoDI, říjen 2001 [on-line] 2001. [cit. 2004-12-02] Dostupné na Internetu: (http://jodi.ecs.soton.ac.uk/noticeboard/wayback.html).

Kahle B., *The Internet Archive: Editors' Interview* [on-line] RLG DigiNews, June 15, 2002, vol 6, num 3, [cit. 2004-12-02] Dostupné na Internetu: (http://www.rlg.org/legacy/preserv/diginews/diginews6-3.html). ISSN 1093-5371.

Lyman P., *Archiving The World Wide Web: Implications for Long-term Preservation*, LOOP: AIGA Journal of Interaction Design Education, num 7 [on-line] 2003. [cit. 2004-12-02] Dostupné na Internetu: (http://loop1.aiga.org/content.cfm?ContentID=99).

Notess G. R., *The Wayback Machine: The Webs Archive*, Online mag, vol 26, num 2 [on-line] 2002. [cit. 2004-12-02] Dostupné na Internetu: (http://www.onlinemag.net/mar02/OnTheNet.htm).

Tybursky G., *The Internet Archive and the Search for Integrity*, [on-line] 2004. [cit. 2004-12-02] Dostupné na Internetu: (http://www.virtualchase.com/articles/internet_archive.html).

Wikipedia*Internet Archive*, Wikipedia [on-line] 2004. [cit 2004-12-02] Dostupné na Internetu: (http://en.wikipedia.org/wiki/Internet_Archive).

Žabička P., *Archivace webu*, [on-line] 2002. [cit. 2004-10-09] Dostupné na Internetu: (http://www.ics.muni.cz/mba/dl02/dlfi02-9.pdf).

Metadata DC

DC.IDENTIFIER = http://www.fi.muni.cz/∼xnovotn8/PV070/esej.pdf : URL
DC.TITLE = Internet Archive
DC.DESCRIPTION = Esej na téma archiv elektronického dědictví Internet Archive.
DC.CREATOR = Luděk Novotný
DC.DATE.Created = 2004-12-02 : ISO8601
DC.DATE.Revsed = 2004-12-12 : ISO8601
DC.LANGUAGE = cs : ISO639-1