Web po 20 letech : co z něj zbude pro budoucí generace?

Ing. Libor Coufal / Národní knihovna České republiky / Libor.Coufal@nkp.cz

Resumé:

Článek rozebírá rychlý vývoj, kterým prošel web od svého vzniku na počátku 90. let. Během té doby se z něho stal důležitý nástroj publikování. Objem online publikování značně vzrostl a přesáhl rozsah tradičního publikování. Materiál publikovaný online představuje významnou část kulturního dědictví. Zároveň je ale velmi nestálý a ohrožený zánikem. Důležitou roli v uchování webu pro budoucí generace sehrávají projekty archivace webu.

Klíčová slova: *web – online publikování – volatilita – crawlery – archivace webu.*

Summary:

The article surveys the rapid development the Web went through since its inception in the early 1990s. During that time, it has developed into an important means of publication. In fact, online publishing grew immensely and its volume has exceeded traditional publishing. The material published online constitutes a significant part of cultural heritage. However, it is very volatile and under a constant threat of dissapearing. Web archiving initiatives play an important role in preserving the Web for future generations.

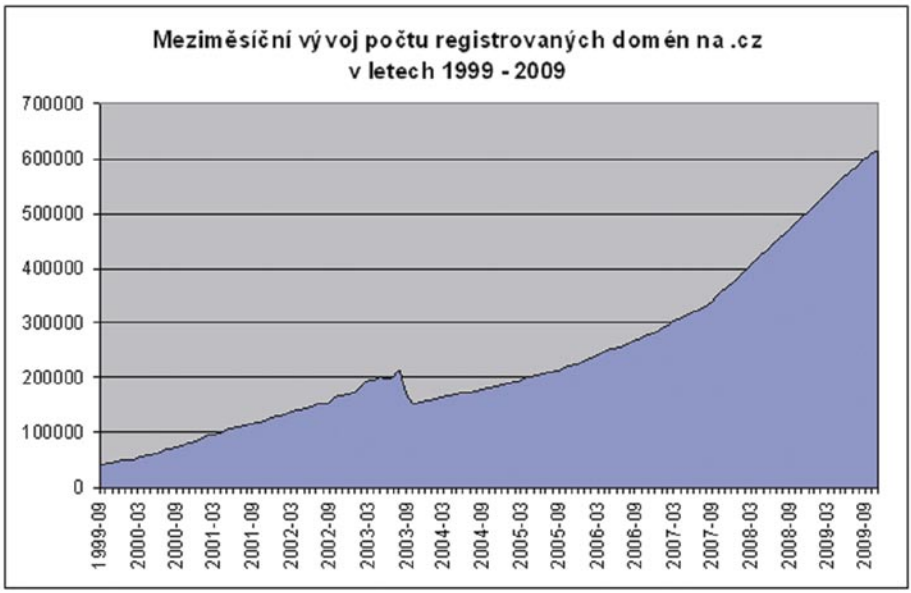
Keywords: *Web – online publishing – volatility – crawlers – web archining.*

Globální informační systém World Wide Web (zkráceně WWW nebo také jen web) letos v březnu oslavil 20. narozeniny (CERN. World...). V roce 1989 Tim Berners-Lee z European Organization for Nuclear Research (CERN) v Bernu navrhl koncept propojení osobních počítačů přes internet pomocí hypertextu. V roce 1990 byl v CERN zprovozněn první prototyp webového serveru a o rok později – konkrétně 6. srpna 1991 ve 14:56:20 hod. – spatřila světlo světa první veřejná webová stránka (Brown, 2006, s. 1). Během pouhých několika let došlo k masivnímu rozšíření webu. Ke konci roku 1993 existovalo jen něco přes 500 známých webových serverů a na WWW připadalo 1 % provozu na internetu. Na konci roku 1994 už to bylo 10 tisíc serverů a 10 milionů uživatelů. (CERN. How...)

Přesnou velikost webu je téměř nemožné zjistit. Existují pouze odhady, které se ovšem značně liší. Studie (Gulli a Signorini, 2005) na základě téměř 440 tis. dotazů v 75 jazycích odhadla velikost ke konci ledna 2005 na 11,5 miliardy webových stránek ve veřejné indexovatelné části webu. Server WorldWideWebSize.com[1] odhaduje velikost indexovatelného (povrchového) webu na začátku listopadu 2009 na nejméně 21,68 miliard stránek. Povrchový web ovšem představuje pouze pomyslný vrcholek ledovce. Jeho skutečná velikost, která zůstává z velké části skrytá pod povrchem, je mnohonásobně větší. Často citovaná studie (Bergman, 2001) odhaduje celkovou velikost webu na 550 miliard stránek, většinu z nich v hlubokém webu.

Známý webový vyhledávač Google oznámil v červenci 2008, že jeho roboti překročili při procházení webu další milník – 1 bilion jedinečných URL (Alpert, 2008). Toto číslo je ovšem třeba brát s velkou rezervou, neboť ne každá jedinečná URL je zároveň smysluplná – velká část těchto stránek je automaticky generována pomocí CMS, webovými kalendáři apod. Jedná se tedy o obsah, který z informačního hlediska nemá velkou hodnotu.

Poměrně přesné statistiky existují o počtu registrovaných domén. Na začátku listopadu 2009 bylo na světě registrováno celkem více než 112,6 milionů aktivních domén, z toho nejvíce – cca 83 mil. – na TLD.com (DomainTools, 2009). V České republice bylo v září 1999 v rámci TLD.cz registrovaných 41 570 domén, zatímco v listopadu 2009 to bylo už více než 612 tisíc domén (viz obr. 1). Jen během posledního roku došlo k nárůstu o téměř 100 tisíc (cca 19,5 %).[2]



Z uvedených statistik je patrné, že se web během velmi krátké doby rozrostl do obřích rozměrů – řádově stovek terabytů až několik petabytů – a jeho velikost neustále roste. Podle Lymana s Varianem (2003, s. 2) „World Wide Web obsahuje kolem 170 terabytů informací na svém povrchu; to je sedmnáctkrát větší objem, než velikost tištěných fondů Kongresové knihovny.“ I pokud vezmeme v potaz netextový (fotografický a filmový) materiál ve fondech knihoven, můžeme rozumně předpokládat, že objem informací, které obsahuje web, je stejně velký, nebo dokonce větší, než v největších světových knihovnách (Masanès, 2006d, s. 78–79).

Web jako publikační médium

Web je hlavní publikační aplikace internetu. Je to jedinečný informační systém, který lze využít pro generování, aktualizování a publikování obsahu všemi představitelnými způsoby, které nabízí moderní výpočetní technika. (Masanès, 2006c, s. 15) S jeho vznikem se tak objevil snadno dostupný, všudypřítomný a relativně levný prostředek pro tvorbu a šíření informací. Ve srovnání s tradičními publikačními médii jde o revoluci v publikování, která posunuje hranice možností ve všech směrech. „Tak jako tiskařský lis stimuloval moderní vydavatelský průmysl, technologie a všudypřítomnost webu znamenaly převrat v šíření intelektuálního vlastnictví“ (Iowa State University).

Díky webu mohou organizace i jednotlivci velmi snadno a s minimálními náklady zprostředkovat zajímavé a aktuální informace potenciálně velkému okruhu zájemců, ať už pomocí klasických webových stránek nebo velmi populárních blogů.[3] V souvislosti s tím došlo k přesunu značné části publikování do online podoby, což vedlo k masivnímu

nárůstu publikovaných informací a zejména pak k radikální změně způsobu, jakým jsou informace používány.

Vývoj webu dramaticky zvýšil nejen množství publikovaných informací, ale také počet potenciálních vydavatelů. V byznysu publikování online se může vydavatelem stát téměř kdokoliv. Výstižně to vyjádřil autor blogu Dosh Dosh, zaměřeného na blogování a internetový marketing (Maki, 2008): „Díky extrémně nízkým bariérám vstupu a variabilním nákladům umožňuje web komukoliv s počítačem stát se nezávislým vydavatelem: výsledkem je, že množství a rozmanitost online obsahu ve většině oborů daleko přesahuje tištěné publikace."

Web byl ovšem od samého počátku zamýšlen nejen jako zdroj informací. Uživatelé se také měli aktivně podílet na vytváření jeho obsahu. První webový browser, který vyvinul Berners-Lee, sloužil nejen k prohlížení, ale zároveň i k publikování. V tomto světle je třeba vnímat nejen masivní rozvoj online publikování, ale také vznik tzv. webu 2.0 s velkým podílem obsahu generovaného uživateli.

Publikování online nepřináší jen samá pozitiva: ve srovnání s klasickým způsobem publikování má minimálně jeden velký nedostatek. Tradiční tištěné publikace jsou po vydání nezávislé na svých vydavatelích. Naproti tomu webové publikace existují pouze na serverech svých tvůrců, a proto jsou závislé na permanentním publikování těmito tvůrci (Masanès, 2006c, s. 15). Jedním z řešení, jak odstranit tuto závislost, může být archivace webu.

Jak jsme ilustrovali v předchozí části, objem informací publikovaných na webu neustále roste. Nejde přitom jen o informace vědeckého charakteru, ale také o osobní výpovědi dokumentující běžný život. Velká část těchto informací není dostupná v jiné formě a zároveň má krátký poločas rozpadu, takže existuje riziko jejich nenávratné ztráty. Přitom jde o nedílnou součást historického dědictví a důležitý zdroj poznání pro budoucí generace.

Pomíjivost webu

Webový obsah je velmi volatilní. Prakticky každý návštěvník webu má bohaté zkušenosti s chybovým hlášením 404 – page not found (stránka nenalezena). Snadnost, s jakou lze na webu publikovat, sebou bohužel přináší také časté aktualizace a změny, kdy je původní obsah nahrazen novým a zmizí. Většinou se jedná o změnu jednotlivých stránek nebo jejich částí, ale v extrémních případech dochází k zániku celých webů.

Ne vždy dochází k úplnému zániku obsahu stránek a trvalé ztrátě informací. Často jde pouze o přesun na jinou URL adresu, ať už v rámci původní domény nebo na jinou doménu. Jedná se o problém trvalosti odkazů – v angličtině se pro tyto případy používá termín „zahnívání" odkazů (link rot). V takovém případě je častokrát možné obsah dohledat pomocí vyhledávačů či jinými prostředky. Např. studie (Lawrence et al., 2001) odhalila 53 % neplatných odkazů po šesti letech od publikování. Většina z těchto dokumentů však byla stále dostupných na jiných URL a bylo možno nalézt buď originální dokument, nebo alespoň související informace. Pouhá 3 % dokumentů se nepodařilo objevit vůbec. Několik dalších podobně zaměřených studií pro lékařské a vědecké časopisy také dochází k podobným závěrům, že volatilita odkazů je větší než volatilita samotného obsahu (Dellavalle et al., 2003; Wren, 2008). Je tedy nutno rozlišovat mezi volatilitou webu a volatilitou URL.

Řada odhadů se pokouší o stanovení průměrné životnosti webové stránky nebo URL. Zakladatel Internet Archive Brewster Kahle např. uvádí průměrnou životnost URL 44 dnů (Kahle, 1997) a průměrnou životnost webové stránky 100 dní (Kornblum, 2001). Studie Kalifornské univerzity v Berkeley (Lyman a Varian, 2000) cituje[4] průměrnou životnost webové stránky 44 dnů, zatímco (Lawrence et al., 2001, s. 30) ji odhadují na 75 dnů. Z praktického hlediska ovšem taková čísla nemají příliš velkou vypovídací hodnotu. Je velmi sporné pokoušet se určit jednu průměrnou hodnotu, protože ta se může značně lišit v závislosti na typu stránek. Velké firmy nebo státní instituce mají často propracované procesy údržby a archivace webových stránek, a proto tento typ obsahu bývá

poměrně stabilní. Na opačném konci spektra leží stránky, které slouží pro účely dokumentace určité aktuální, většinou krátkodobé, události. Tyto stránky jsou naopak velmi náchylné k častým změnám a mívají krátkou životnost. Mezi těmito dvěma extrémy leží celé spektrum různých typů webů s různou délkou životnosti.

Některé typy webových zdrojů jsou pravděpodobně více stabilní než jiné, např. můžeme poměrně logicky předpokládat, že multimediální obsah bude stabilnější než textový obsah. Přesto zůstává faktem, že webový obsah jako celek je, nebo přinejmenším má velký potenciál být, pomíjivý. Proto je třeba konat, abychom eliminovali nebo alespoň snížili rizika, která tato pomíjivost přináší. Ne vše, co kdy bylo publikováno na webu, má stejnou hodnotu. Z praktických důvodů je zapotřebí učinit rozhodnutí o tom, který materiál je hodnotný a měl by být uchován. Je také třeba snažit se pochopit, jaké faktory ovlivňují volatilitu a tím pádem riziko zmizení obsahu.

Archivace webu

Masanès (2006c, s. 7–8) tvrdí, že: „Web není a nikdy nebude médiem, které by bylo schopno se samo uchovávat. Jednu ze základních příčin je třeba hledat v rozporu mezi publikováním a uchováváním. Publikování představuje vytváření nového, a to i na úkor starého... Zkušenosti ukazují, že samotní tvůrci obsahu nemají dostatečnou motivaci k uchovávání, aby se na ně dalo v tomto ohledu spoléhat. Prvním krokem k uchovávání totiž je, aby bylo prováděno organizací jiného druhu, vedenou jinými cíli, pohnutkami a dokonce i jinou etikou. Web, jako informační infrastruktura, není schopen vyřešit problém, který je převážně organizačního původu. Proto je potřebná archivace webu, jako aktivita nezávislá na publikování.“

Web je velmi rozmanitý a zájem na jeho uchovávání může být u různých typů organizací veden odlišnými motivy. Např. archivy mají zájem o ty stránky, které obsahují materiály archivní povahy, zatímco knihovny se zaměřují spíše na publikace a jiné zdroje, zajímavé z hlediska jejich uživatelů. Podle Daye (2003) tedy není jasné, kdo by měl být za archivaci webu zodpovědný. Globální povaha webu také značné komplikuje vymezení „národních hranic“ zodpovědnosti za jeho uchovávání.

První webové archivy se začaly objevovat záhy po vzniku webu, v polovině 90. let minulého století. Nejčastěji se archivací webu zabývají národní knihovny. Za průkopníky na tomto poli jsou považovány tři instituce: Internet Archive, Národní knihovna Švédska (Kungliga biblioteket) a Národní knihovna Austrálie (National Library of Australia), které začaly své projekty zhruba ve stejné době v rozmezí let 1995 – 1996. V roce 2003 bylo založeno mezinárodní sdružení International Internet Preservation Consortium (IIPC), které sdružuje webové archivy z více než 30 zemí Evropy, Severní Ameriky a Australasie.[5] Nejčastěji se používají dva přístupy k archivaci webu:

- budování výběrových archivů založených na ručním výběru stránek na základě určitých kritérií, nejčastěji jejich dlouhodobé badatelské hodnoty;
- celoplošný přístup založený na automatizovaném sběru velkého množství stránek, nejčastěji v rámci národní domény, pomocí robotů.

Hlavní výhodou automatizovaných plošných sklizní je to, že představují relativně levný způsob sběru velkého množství webového obsahu. Současné technologie pro automatizovaný sběr si ovšem nedokáží poradit s některými rozšířenými technikami tvorby webových stránek. Ukazuje se, že metoda automatizovaného sklízení má značné nedostatky, zejména pokud jde o nefunkční a chybějící obsah včetně odkazů. Manuální kontrola kvality je u tohoto přístupu možná přinejlepším pouze na malém vybraném vzorku. Celoplošný přístup „není náhradou za specializovanou iniciativu sběru a uchovávání webu... Poskytuje však cenný a rozsáhlý obraz, který je možno použít jako doplněk jakékoliv úžeji zaměřené iniciativy“ (Day, 2003, s. 10).

Výběrový přístup sice umožňuje ošetřit některé z těchto problémů, ale vyžaduje vysoký podíl lidské práce a množství zdrojů, které takto mohou být archivovány, je velmi

omezené. Protože je výběrová archivace webu manuálně náročná, je mnohem nákladnější než automatické sklízení – podle některých odhadů dokonce více než stokrát (Day, 2003, s. 23). Výběrové sklizně také umožňují ošetřit práva pro zpřístupnění. To sice na jednu stranu zvyšuje náklady, ale zároveň umožňuje otevření archivu a vede k jeho vyššímu využívání.

Jedním z hlavních rysů webu je vzájemná propojenost jeho obsahu. Důsledkem toho je, že archivace webu v sobě vždy nutně zahrnuje určitý stupeň selektivnosti, i když to nemusí být striktně ve smyslu manuálního výběru jednotlivých stránek. Např. celoplošné sklizně jsou sice výsledkem automatického procesu, ten je ale zásadně ovlivněn rozhodnutími o souboru vstupních URL, rozsahu, frekvenci sklízení, délce trvání, nastavení limitů pro hloubku zanoření a počet stahovaných souborů, apod. (Masanès, 2006d,s. 76).

Jakýkoliv projekt archivace webu by se měl snažit o co nejkomplexnější pojetí. V praxi ovšem i ten nejplošnější přístup může zachytit pouze jakýsi vzorek webu – určitý průřez webem v čase a prostoru. „Jak učinit tento vzorek smysluplný a odrážející celý web? Jaké implikace to bude mít pro budoucí chápání toho, co byl web? Všechny tyto otázky je třeba zvažovat při angažování se v archivaci webu“ (Masanès, 2006c, s. 17).

Rozhodnout, která z těchto metod je lepší, je velmi složité. Každá z nich má své přednosti i nedostatky a obě se navzájem doplňují. Proto webové archivy často volí kombinaci obou přístupů.

Akvizice webového obsahu

Získávání dokumentů, které se v odborné terminologii nazývá akvizice, je tradiční knihovnická činnost. V prostředí webových archivů jde o velmi komplikovaný proces, který sestává ze dvou základních kroků: 1) zkopírování a uložení všech stránek včetně souvisejících obrázků a ostatních souborů a 2) přepsání všech odkazů tak, aby nevedly zpět na web, ale směřovaly na uložené stránky v archivu. Existuje několik možností, jak se s tímto složitým úkolem vypořádat: archivace může probíhat na straně serveru nebo vzdáleně na straně klienta. Ani jedna z nich však sama o sobě nedokáže efektivně pokrýt celý rozsah technik používaných pro publikování na webu.

Archivace na straně serveru může probíhat dvěma způsoby. Transakční archivace sleduje, jak se skuteční uživatelé pohybují v rámci stránek, a zaznamenává jednotlivé transakce na serveru včetně archivace doručeného obsahu. Druhá možnost spočívá v přímém kopírování dokumentů ze serveru bez pomoci HTTP. Tyto metody ovšem mají jedno zásadní omezení: vyžadují autorizaci ze strany vydavatele a jeho aktivní spolupráci. Nelze je tedy automatizovat a musí být řešeny vždy individuálně, proto jsou využitelné pouze pro velmi malé archivy. Pro velké projekty archivace webu nejsou kvůli špatné škálovatelnosti vhodné. Z těchto důvodů nejsou v praxi příliš rozšířené. Omezíme se tedy pouze na konstatování, že mohou být využity spíše doplňkově v těch případech, kdy nelze použít archivaci na straně klienta nebo tam, kde tato metoda neposkytuje uspokojivé výsledky.

Nejpoužívanější metodou je archivace na straně klienta pomocí speciálních robotů nazývaných většinou crawlery.[6] Tito roboti umožňují značnou automatizaci archivace a jsou dobře škálovatelní, proto je tato metoda vhodná i pro velké webové archivy. Na druhou stranu ovšem archivace webu na straně klienta zahrnuje i některá významná omezení.

Hlavní problém archivace na straně klienta spočívá v tom, že webový protokol HTTP neposkytuje (na rozdíl např. od ftp) seznam všech dokumentů na webovém serveru a nedokáže doručit celý jeho obsah v dávkách. HTTP servery jsou schopné doručit pouze jednotlivé soubory v jednorázových transakcích na základě požadované konkrétní URL (Masanès, 2006c, s. 21). Aby crawler mohl archivovat webovou stránku, musí znát její URL nebo ji musí objevit následováním odkazů z jiných známých stránek. Z toho vyplývá významné omezení: aby mohla být stránka nalezena, musí na ni odkazovat jiné stránky.

Crawlery používají pro procházení webu techniku parsování HTML kódu a extrakce odkazů.[7] Na začátku je crawleru zadán seznam URL, tzv. semínek (angl. seeds), ze kte-

rých vychází, a stanoveny hranice prostoru, v rámci kterého se má pohybovat (angl. scope) – např. doména .cz. Crawler postupně zpracovává všechna semínka tím, že stahuje dokumenty z jednotlivých URL. Pro každou stránku analyzuje (parsuje) její HTML kód a extrahuje z něj odkazy na další stránky. Nalezené odkazy přidává k původnímu seznamu URL a tento postup neustále opakuje, dokud nezpracuje celý seznam. Roche (2006) demonstruje vnitřní mechanismus fungování jednoduchého crawleru, určeného pro potřeby malých archivů nebo jednotlivců. Detailní popis crawleru vyvinutého specificky pro účely archivace webu ve velkém objemu je možno nalézt v článku (Mohr et al., 2004).

Každý crawler se obecně skládá ze dvou hlavních komponent. První z nich parsuje HTML kód stránek a hledá v něm odkazy. „Parser" skenuje pouze ty soubory, ve kterých by se potenciálně mohly vyskytovat odkazy, především HTML stránky, a nezkoumá např. obrázkové nebo zvukové soubory. Soustředí se přitom pouze na tu část HTML kódu, která se používá pro odkazování – jde zejména o značky <img> a <a>, resp. jejich atributy src a href.[8] Zbytek kódu stejně jako vlastní textový obsah ignoruje. Parser musí udržovat seznam všech zpracovaných odkazů, aby se nezpracovávaly duplicitně. Nalezené odkazy je třeba porovnat vůči definici vymezeného prostoru (scope). Odkazy, které vedou mimo tento parametr, nejsou relevantní a dokumenty na nich se nearchivují.[9] Odkazy, které této definici vyhovují, jsou sesbírány a zařazeny do seznamu semínek, kde čekají na další zpracování. Druhý z modulů crawleru postupně zpracovává semínka – připojuje se k jednotlivým serverům, odesílá požadavky a řídí stahování souborů. Stažené soubory, které potenciálně mohou obsahovat odkazy, posílá zpět na parser ke skenování a extrakci odkazů. Ostatní soubory (např. obrázky nebo videa) pouze uloží na disk.

Po stáhnutí a uložení stránek do archivu je potřeba změnit odkazy v archivovaných dokumentech tak, aby vedly do archivu. V opačném případě by byl uživatel při prohlížení archivovaných verzí neustále odkazován zpět na živý web. Prakticky je zapotřebí změnit všechny absolutní odkazy na relativní. Původní relativní odkazy většinou mohou zůstat v nezměněné podobě. Pokud jsou ale tvořeny nestandardní cestou, např. pomocí definice výchozího adresáře v hlavičce stránky, nemusí tyto odkazy fungovat správně.

Změnu odkazů lze provést několika způsoby. Jednou z možností je trvalé přepsání odkazů přímo v kódu stránek, což je poměrně snadné uvnitř HTML, ale uvnitř komplexních skriptů je to téměř neřešitelný problém (Roche, 2006, s. 97–98). Druhou možností je ponechat odkazy v původní podobě a změnu provádět dynamicky, „na vyžádání", při prohlížení stránek v archivu, např. pomocí java scriptu (Masanès, 2006c, s. 34).

Výkon crawleru je optimalizován tím, že oba moduly pracují paralelně. Crawler také může být připojen a stahovat najednou z více serverů. Díky vysoké výkonnosti by při rychlých přenosových linkách crawler mohl snadno způsobit přetížení serveru. Proto je třeba nastavit prodlení mezi jednotlivými požadavky a tím omezit počet odesílaných požadavků za jednotku času. K „dobrému vychování" robotů patří také respektování pokynů webmasterů v souborech robots.txt, které umožňují vyloučit některé stránky nebo jejich části z procházení a indexace.[10]

Crawlery jsou velmi efektivní v tom smyslu, že dokáží prozkoumat velkou část webu a pomocí odkazů najít nové stránky, i když začínají z malé počáteční sady semínek. Jak ale bylo uvedeno výše, mají také řadu vážných omezení. Většina z nich se vztahuje k problémům s extrakcí odkazů nebo se stahováním obsahu. Boyko (2004) sestavil taxonomii a charakteristiku nejčastějších typů problémů, se kterými se může setkat crawler při procházení a archivaci webu. Mezi poměrně častá omezení, která mohou potenciálně způsobit crawlerům problémy, patří např. nesprávně formované odkazy, komplexní odkazy obsahující parametry, odkazy tvořené pomocí skriptovacích jazyků, přesměrování, nutnost autorizace nebo jiný protokol než HTTP.

V rámci webu dochází k permanentnímu vývoji nových, inovativních technologií. Moderní webové stránky často používají komplexní java skript, AJAX nebo vložené binární soubory, vytvořené v Java nebo Flash. V souvislosti s tím narůstá jejich komplexnost a sofistikovanost, takže je stále složitější je archivovat. Přestože moderní crawlery neustále na tento vývoj reagují a jejich schopnosti se s každou novou verzí lepší, mají v sobě díky používaným technologiím, založeným na extrakci odkazů, zabudovány určité inherentní hranice, které nejsou schopny překročit. Archivace na straně klienta pomocí

extrakce odkazů si neumí poradit s hlubokým webem, streamovanými technologiemi nebo stránkami, vyžadujícími interakci s uživatelem. „Prostá evoluce již není dostačující a je třeba vyvinout nové, revoluční přístupy ke sklízení webu." (Celbová et al., 2008, s. 20).

Archivace hlubokého webu

Crawlery fungují na principu „nalézání cesty" (path finding). Každá stránka, která má být archivována, musí být nejprve objevena, tzn. musí k ní vést viditelná stezka z jiné webové stránky. Z tohoto důvodu je velká část webu pro crawlery neviditelná a nemůže jimi být dosažena a tudíž ani archivována. Tato část webu je nazývána hluboký nebo skrytý web.

Nedostupnost hlubokého webu pro crawlery je čistě technický problém. Hranice mezi tím, co je a není dostupné, se neustále mění a posunuje oběma směry v závislosti na tom, jak se vyvíjejí technologie webu i crawlerů. Překotný vývoj webových technologií na jedné straně neustále vytváří pro crawlery nové výzvy a problémy, na druhé straně ale zároveň stimuluje jejich další vývoj a tím zlepšuje jejich schopnosti. Např. weby používající pro navigaci Flash byly dříve pro crawlery nedostupné. Uvolnění specifikace Flash ale umožnilo extrakci odkazů z těchto stránek, takže crawlery jsou nyní schopny procházet a tyto stránky už nejsou součástí hlubokého webu (Masanès, 2006a, s. 115–116). Je jasné, že technologie crawlerů budou vždy pokulhávat za vývojem webu.

Jednou z nejčastějších příčin nedostupnosti hlubokého webu je nevhodné používání databází. Stránky, které jsou generovány dynamicky pomocí databáze, nepředstavují samy o sobě problém. Důležité je to, že výstupem musí být HTML stránka s jasně definovanou cestou pomocí URL, aby jí crawler byl schopen najít. Problém vyvstává v momentě, kdy pro vygenerování HTML a URL je zapotřebí, aby uživatel např. na něco kliknul, vybral položku z menu nebo specifikoval dotaz a spustil vyhledávání. Tyto případy, které vyžadují interakci reálných uživatelů, lze jen obtížně automatizovat.

Zpřístupnění webových archivů

Uchovávání a zpřístupnění dokumentů jdou tradičně ruku v ruce: aby vůbec mohly být zpřístupněny, musí být nejprve uchovány a obráceně, pokud nebudou nikdy zpřístupněny, nemá smysl ani jejich uchovávání. Lze tedy říct, že bez zpřístupnění není uchovávání a naopak. To platí obecně pro všechny typy dokumentů, ale pro webové dokumenty dvojnásob. Většina obsahu na webu je volně dostupná a uživatelé očekávají totéž také pro webové archivy. „Většina současných aktivit spojených s internetovými archivy se zaměřuje na shromažďování a uchovávání informací. Úspěch internetu je ale založen na snadném přístupu k informacím. Je proto rozumné očekávat, že se výsledný úspěch jakéhokoliv internetového archivu bude měřit na základě prostředků, které tento archiv bude poskytovat pro přístup k uchovávaným materiálům" (IIPC, 2006, s. 1). Webové archivy bez přístupu, tzv. černé nebo tmavé archivy, nemohou uživatelům nabídnout očekávanou hodnotu.

Na druhé straně, některé webové stránky přinášejí svým vydavatelům příjmy z reklamy nebo předplatného a webové archivy si musí najít takovou pozici, ve které nebudou pro originální stránky představovat konkurenci. Toho lze dosáhnout např. respektováním omezení v souborech robots.txt, zpřístupněním archivu s určitým zpožděním, omezením funkcionality nebo nižší rychlostí připojení (Masanès, 2006c, s. 9). Hlavní využití webových archivů by ale mělo spočívat zejména v poskytování přístupu k obsahu, který již není na webu k dispozici, což případné příjmy vydavatele nijak neohrožuje.

Možnosti zpřístupnění webových archivů závisí na legislativě týkající se autorských práv a povinného výtisku, která se v jednotlivých zemích může značně lišit. Jde o komplexní problematiku, která je nad rámec tohoto článku. Případným zájemcům o tuto problematiku doporučujeme k získání podrobnějších informací např. studie (Celbová, 2008 nebo Charlesworth, 2003).

Uživatelé jsou z webu zvyklí na určitý komfort – většina obsahu je přístupná bez omezení, z jakéhokoliv počítače připojeného k internetu, takže je v podstatě vzdálená

pouze několik kliknutí myší. Běžným standardem je vyhledávání v plných textech milionů dokumentů a výsledky řazené podle relevance. Vyhledávání by mělo být uživatelsky přívětivé a co nejméně komplikované. Uživatelé nechtějí používat logické a proximitní operátory, řízené slovníky, předmětová hesla nebo vyhledávání v různých indexech. Zejména nejpopulárnější webový vyhledávač Google se svým jednoduchým, až minimalistickým, rozhraním znamenal revoluci ve vyhledávání. Od nástupu Googlu se stal standardem jednoduchý vyhledávací řádek, kam stačí napsat pouze několik klíčových slov.[11]

Většina webových archivů v současnosti umožňuje pouze vyhledávání pomocí URL, což předpokládá, že uživatel dopředu zná konkrétní URL adresu hledaného zdroje. To je z pohledu uživatelů nedostatečné, a proto je zapotřebí indexovat archivy pro vyhledávání v plných textech s odkazy na plné texty archivovaných dokumentů ve výsledcích. Webové archivy ovšem pracují s extrémně velkými objemy dat v řádech stovek miliónů až miliard dokumentů. Fulltextová indexace v těchto objemech je velmi komplexní úkol, který vyžaduje značné výpočetní kapacity, a výsledný index webového archivu může být extrémně velký. Použité aplikace tedy musí být škálovatelné a schopné poradit si s tímto objemem dat.

Důležitou vlastností webových archivů z hlediska zpřístupňování je jejich časová dimenze. Zatímco na webu existuje vždy pouze jediná, aktuální verze stránek, v archivech může být jedna stránka zastoupena mnohonásobně v několika verzích z různých dat (tj. archivovaných v různých obdobích). Archivace probíhá opakovaně prostřednictvím „otisků“ webu, sejmutých v určitém čase, a archivované dokumenty jsou proto vždy jen konkrétní verzí z konkrétního data. Jednotlivé časové verze dokumentů mohou být shodné (duplikáty) nebo se mohou lišit. Navíc může při brouzdání v archivu docházet ke skokům v čase, protože dokumenty z určitého data, které nejsou k dispozici, mohou být automaticky nahrazeny nejbližší dostupnou časovou verzí. Proto je důležité při zpřístupnění webových archivů tento časový aspekt ošetřit a patřičně jej zdůraznit. U každého dokumentu je třeba zachytit, kdy byl archivován, a toto datum zobrazit při prohlížení, aby bylo uživatelům vždy patrné, kde přesně z hlediska času se v archivu právě nacházejí.

Webové archivy vyžadují sofistikované vyhledávací technologie: kvalitní vyhledávače a vhodné prohlížecí rozhraní, které společně dokáží nejen vyhledávat v indexech a zobrazovat výsledky, ale také pracovat s prohlížením časových verzí. Množství prohledávaných dokumentů, a tím pádem i potenciálních výsledků je příliš velké, proto musí být výsledky vyhledávání ohodnoceny a smysluplně setříděny na základě relevance nalezených dokumentů. Zobrazovací rozhraní by také mělo podporovat navigaci v rámci archivu dynamickou úpravou odkazů tak, aby směřovaly do archivu a ne zpět na živý web.

Závěr

Web není jen pouhým obrazem moderní společnosti, ale sám se podílí na jejím formování. Webové archivy se snaží zachovat co největší část tohoto bohatého informačního zdroje pro budoucí generace. Ne vše, co je publikováno na webu, lze archivovat pomocí současných nástrojů. S tím, jak se webové technologie neustále rozvíjejí, narážejí tyto nástroje na hranice svých možností. Je potřeba vyvinout zcela nové nástroje, založené na jiných premisách. Archivace je ovšem pouze prvním krokem na cestě k uchování webu. Z dlouhodobého hlediska bude stále významnější roli hrát zajištění trvalého přístupu k archivovaným materiálům.

Použitá literatura:

ALPERT, Jesse; HAJAJ, Nissan. We knew the Web was big... *The official Google blog* [online]. 2008 [cit. 2009-11-25]. Dostupný z WWW: <http://googleblog.blogspot.com/2008/07/we-knew¬-web-was-big.html>.

BERGMAN, Michael K. *The deep web : surfacing hidden value*. Sioux Falls : BrightPlanet [online]. 2001 [cit. 2009-11-25]. Dostupný z WWW: <http://www.brightplanet.com/images/uploads/De-e¬pWebWhitePaper_20091015.pdf>.

BOYKO, Andrew. *Test bed taxonomy for trawler* [online]. 2004 [cit. 2009-11-25]. Dostupný z WWW: <www.netpreserve.org/publications/iipc-r-002.pdf>.

BROWN, Adrian. *Archiving websites : a practical guide for information management professionals.* Londýn : Facet, 2006. 238 s. ISBN 9781856045537.

CELBOVÁ, Ludmila. Český web a povinný výtisk – jde to spolu dohromady? *Knihovna* [online]. 2008, roč. 19, č. 2, s. 59-75 [cit. 2009-11-25]. Dostupný z WWW: <http://knihovna.nkp.cz/kniho-v¬na82/82005.htm>.

CELBOVÁ, Ludmila. *Archivace webu*. Praha : Národní knihovna ČR, 2008. 45 s. ISBN 9788070505625.

CERN. *How the web began* [online]. 2008 [cit. 2009-11-25]. Dostupný z WWW: <http://public.web.cern.ch/public/en/About/WebStory-en.html>.

CERN. *World Wide Web @ 20* [online]. 2009 [cit. 2009-11-25]. Dostupný z WWW: <http://info.cern.ch/www20>.

DAY, Michael. *Collecting and preserving the World Wide Web* [online]. Bristol : JISC, 2003 [cit. 2009-11-25]. Dostupný z WWW: <http://www.jisc.ac.uk/media/documents/programmes/preservation/archiving_feasibility.pdf>.

DELLAVALLE, Robert P.; HESTER, Eric J.; HEILIG, Lauren F.; DRAKE, Amanda L.; KUNTZMAN, Jeff W.; GRABER, Marla; SCHILLING, Lisa M. Going, Going, Gone : Lost Internet References. *Science*. 2003, vol. 302, no. 5646, s. 787-788. ISSN 0036-8075 (print), 1095-9203 (online). DOI 10.1126/science.1088234.

Domain counts & Internet statistics [online]. 2009 [cit. 2009-11-25]. Dostupný z WWW: <http://www.domaintools.com/internet-statistics>.

GULLI, Antonio; SIGNORINI, Alessio. *The indexable Web is more than 11.5 billion pages* [on line]. 2005 [cit. 2009-11-25]. Dostupný z WWW: <http://www.cs.uiowa.edu/~asignori/papers/the-in-dexable-web-is-more-than-11.5-billion-pages>..

CHARLESWORTH, Andrew. *Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia* [online]. Bristol : JISC, 2003 [cit. 2009-11-25]. Dostupný z WWW: <http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf>.

Use cases for access to Internet archives. In *IIPC Access Working Group* [online]. 2006 [cit. 2009-11-25]. Dostupný z WWW: <http://www.netpreserve.org/publications/reports.php?id=003>.

Guide 3: open access versus traditional publishing [online]. Iowa State University, 2006 [cit. 2009-11-25]. Dostupný z WWW: <http://www.grad-college.iastate.edu/thesis/documents/UMI_Open%20Access.pdf>.

KAHLE, Brewster. *Preserving the Internet* [online]. Scientific American. 1997, vol. 276, no. 3, s. 72-73. ISSN 0036-8733. [cit. 2009-11-25]. Dostupný z WWW: <http://web.archive.org/web/19980627072808/http://www.sciam.com/0397issue/0397kahle.html>.

KORNBLUM, Janet. Web-page database goes Wayback when [online]. In *USA Today*. 2001-10-30 [cit. 2009-11-25]. Dostupný z WWW: <http://www.usatoday.com/life/cyber/tech/2001/10/30/ebrief.htm>.

LAWRENCE, Steve; PENNOCK, David; FLAKE, Garry; KROVETZ, Bob; COETZEE, Frans; GLOVER, Eric; NIELSEN, Finn; KRUGER, Andries; GILES, Lee. Persistence of Web References in ScientificResearch [online]. *IEEE Computer*. 2001, vol. 34, no. 2, s. 26-31 [cit. 2009-11-25]. Dostupný z WWW: <http://www.searchlores.org/library/persi-stence-computer01.pdf>.

LYMAN, Peter; VARIAN, Hal R. *How much information?* [online] University of Kalifornia, 2000 [cit. 2009-11-25]. Dostupný z WWW: <http://www2.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf>.

MAKI. The future of content in the age of information overload [online]. *Dosh Dosh*. 2008 [cit. 2009-11-25]. Dostupný z WWW: <http://www.doshdosh.com/future-of-content-in-the-age-of-information-overload>.

MASANÈS, Julien. Archiving the hidden web. In Julien Masanès (ed.). *Web archiving*. Berlín : Springer, 2006a. 234 s. ISBN: 978-3-540-23338-1.

MASANÈS, Julien. Web archiving. In Marylin Deegan, Simmon Tanner (eds.). *Digital preservation*. Londýn : Facet, 2006b. 260 s. ISBN 978-1-856-04485-1.

MASANÈS, Julien. Web archiving : issues and methods. In Julien Masanès (ed.). *Web archiving*. Berlín : Springer, 2006c. 234 s. ISBN 978-3-540-23338-1.

MASANÈS, Julien. Selection for web archives. In Julien Masanès (ed.). *Web archiving*. Berlín : Springer, 2006d. 234 s. ISBN 978-3-540-23338-1.

MOHR, Gordon; STACK, Michael; RANITOVIC, Igor; AVERY, Dan; KIMPTON, Michele. *An introduction to Heritrix : an open source archival quality web crawler* [online]. IWAW, 2004 [cit. 2009-11-25]. Dostupný z WWW: <http://iwaw.europarchive.org/04/Mohr.pdf>.

ROCHE, Xavier. Copying websites. In Julien Masanès (ed.). *Web archiving*. Berlín : Springer, 2006. 234 s. ISBN 978-3-540-23338-1.

VLČEK, Ivan. *Identification and archiving of the Czech Web outside the national domain* [online]. IWAW, 2008 [cit. 2009-11-25]. Dostupný z WWW: <http://iwaw.europarchive.org/08/IWAW2008-Vlcek.pdf>.

WREN, Jonathan D. URL Decay in MEDLINE - a 4-year Follow-up Study. *Bioinformatics*. 2008, vol. 24, no. 11, s. 1381-85. ISSN 1367-4803.

[1] *http://www.worldwidewebsize.com/*
[2] *http://www.nic.cz/stats/*
[3] *Z anglického web logs. Jde o online deníky, publikované prostřednictvím speciálního publikačního softwaru, který je dostupný buď zdarma, nebo za velmi nízkou cenu, a nevyžaduje znalost tvorby HTML. To usnadňuje publikování i „laikům" a ještě více podporuje přesun publikování online a růst množství publikovaných dokumentů na webu.*
[4] *Je velmi ilustrativní, že původní zdroj této citace – „Size of the Web : A Dynamic Essay for a Dynamic Medium" http://censorware.org/web_size/ - již není dostupný.*
[5] *http://www.nic.cz/stats/*
[6] *Protože tito roboti doslova prolézají web – anglicky crawl. Někdy se místo archivace webu používá termín sklízení webu (angl. web harvesting) a roboti se pak nazývají harvestery.*
[7] *Tato metoda byla původně vyvinuta pro webové vyhledávače a posléze adaptována pro potřeby webových archivů.*
[8] *Odkazy může obsahovat také např. java skript nebo Flash.*
[9] *Mohou být ale použity jako semínka pro další sklizně, např. Národní knihovna ČR je používá pro automatizované sklízení mimo doménu .cz (Vlček, 2008).*
[10] *Některé části stránek mohou být z hlediska robotů problematické, např. mohou vytvořit nekonečné cykly, tzv. pasti (angl. crawler traps). V jiných případech nemusí být procházení, indexace nebo stahování některých částí stránek roboty žádoucí např. kvůli obsahu stránek. V případě výběrových archivací stránek, které jsou ošetřeny dohodou s jejich vydavateli, se tato pravidla ovšem někdy záměrně nerespektují.*
[11] *Pro tento způsob pohodlného vyhledávání se dokonce vžil nový termín „vygůglovat" (z angl. to google).*