WebArchiv – digitální knihovna českého webu

Petr Žabička, MZK Brno a FI MU

S prudkým nárůstem objemu informací publikovaných výhradně na Internetu se úkolem moderní depozitní knihovny stává také shromažďování, ochrana a zpřístupnění online dostupných elektronických informačních zdrojů. V souladu se svým posláním se touto cestou vydala i Národní knihovna ČR, která ve spolupráci s Ústavem výpočetní techniky MU připravuje archiv českého webu.

1 Archivace webu – situace ve světě

V posledních letech exponenciálně roste objem informací dostupných pouze v elektronické podobě na Internetu. Tyto informace jsou však často velmi „křehké“ povahy – velmi rychle se mění, značná část informace po čase nenávratně zaniká (některé studie uvádí, že průměrná životnost webových stránek je asi 100 dnů). Hrozí tak reálné nebezpečí, že v důsledku přechodu na elektronickou formu publikování bez náležitého zajištění archivačních funkcí (jak je známe například v podání klasických knihoven) nebude velká část dnešních informací zachována pro budoucnost, a že budoucí generace budou jednou pohlížet na naši současnost jako na dobu „digitálního temna“. Většina národních knihoven a dalších „paměťových“ institucí, usilujících o uchování kulturního odkazu dané společnosti, hledá proto cesty k tomu, jak rozšířit své tradiční archivační funkce v oblasti tištěné informace i do oblasti informací digitálních. V popředí zájmu jsou přitom zejména informace vznikající na webu.

Jedním z průkopníků na poli archivace webu je americká nezisková organizace *Internet Archive* (`www.archive.org`), jejíž Internetový archiv sahá až do roku 1996 a obsahuje v současnosti přes 160 TB dat. Tato organizace se ve spolupráci s dalšími institucemi snaží (vcelku úspěšně) budovat co nejrozsáhlejší archiv světového webu. Takový záměr je však finančně vysoce nákladný; v letošním roce proto zahájil Internet Archive spolupráci s největšími světovými národními knihovnami s cílem vyvinout novou generaci nástrojů pro archivaci a zpřístupnění webových informačních zdrojů. V průběhu tří let bude na vývoj těchto nástrojů a na archivaci webů zemí zúčastněných knihoven vynaloženo přibližně 3 milióny dolarů. Předpokládá se, že softwarové nástroje vyvinuté v rámci tohoto projektu budou dány k dispozici i ostatním knihovnám pod nějakým typem licence zajišťující volný přístup ke zdrojovým kódům.

Je zřejmé, že žádná knihovna nemá prostředky na to, aby si sama vytvářela archiv celosvětového webu; nemůže přitom ani spoléhat na to, že o dlouhodobou archivaci se postarají vydavatelé elektronických informačních zdrojů. Je proto logické, že se každá vyspělá země snaží (většinou prostřednictvím své národní knihovny) vybudovat alespoň národní archiv elektronických informačních zdrojů – v tomto duchu se nese i připravovaná charta UNESCO o ochraně digitálního kulturního dědictví.

Přístup jednotlivých knihoven k řešení této problematiky se ovšem velmi liší. Některé knihovny, jako například Australská národní knihovna, se snaží archivovat zdroje *výběrově,* tj. zajímají se jen o ty webové zdroje, jejichž kvalitu předem zhodnotí knihovník – viz projekt `pandora.nla.gov.au`. Díky tomuto selektivnímu přístupu čítá archiv australského webu po několika letech provozu pouhých 3675 webových sídel nebo jejich částí, nicméně jedná se (doufejme) o výběr toho „nejdůležitějšího“, co bylo v dané době na webu publikováno. Tento přístup je však velmi náročný na lidské kapacity a proto se většina knihoven vydala jinou cestou: cestou automatizované *plošné archivace* všech dokumentů, které splňují automaticky vyhodnotitelná kritéria. K tomu využívá nejčastěji softwarové nástroje vyvinuté v nejrůznějších projektech v minulých letech (například v projektech severských evropských zemí). Vznikají však i další iniciativy: například ve výše zmíněném konsorciu Internet Archive se po několikaletém zkoumání problematiky rozhodly spojit své síly americká Kongresová knihovna, Britská knihovna, Francouzská národní knihovna a některé severské národní knihovny. Pozadu nezůstává ani Japonská národní knihovna a zahájen byl i projekt na archivaci webových zdrojů v čínštině.

Podobným směrem se v roce 2000 vydala i Národní knihovna ČR, když ve dvouletém pilotním projektu „*Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet*“ zprovoznila za grantové podpory Ministerstva kultury a ve spolupráci s Ústavem výpočetní techniky Masarykovy univerzity infrastrukturu pro tvorbu digitálního archivu českého webu (`webarchiv.nkp.cz`). Práce na archivaci českého webu pokračují i po ukončení pilotního projektu.

2 Projekt WebArchiv

Cílem projektu WebArchiv je *zajištění trvalého uchování domácích elektronických online publikovaných informačních zdrojů jako součásti národního kulturního dědictví*. Vzhledem k povaze, rozmanitosti a množství těchto zdrojů je zřejmé, že stanovení podmínek, které musí archivované elektronické zdroje splňovat, významně ovlivní budoucí hodnotu vytvořeného archivu.

2.1 Výběr zdrojů k archivaci

Pokud padla v úvodu tohoto článku zmínka o „online“ publikovaných zdrojích, je nutné upozornit na to, že již rozhodnutí zaměřit se primárně na „webové“ zdroje znamená, že se zaměřujeme jen na jistou podmnožinu všech existujících online zdrojů. Je zřejmé, že pokus archivovat online elektronické zdroje dostupné jinak než prostřednictvím Internetu by byl velmi nákladný a jeho přínos pro archiv zanedbatelný. Takovéto kategorické tvrzení však již nelze pronést o ne-webových Internetových zdrojích. Většinou totiž nelze dopředu určit, která technologie začne mít v budoucnosti větší význam, a která je jen krátkou epizodou v dějinách Internetu. Přesto lze zatím stále obhájit názor, že většině populace je reálně přístupná jen ta část zdrojů, ke kterým se dostanou prostřednictvím běžného www-prohlížeče a proto právě tato část zdrojů by měla být primárním předmětem zájmu Národní knihovny.

Pokud tedy pomineme relativně velkou množinu mailových a newsových diskusních skupin, zůstává před námi dvojice protokolů http a ftp (protokol gopher lze dnes již považovat za mrtvý, https je určen pro šifrovaný přenos dat a lze jej proto považovat za protokol určený především k přenosu důvěrných informací, které nejsou předmětem veřejného zájmu).

Pokud dosavadní zkušenosti ukazují, že z hlediska dlouhodobé konzervace je opravdu nejvýznamnější část dokumentů dostupná přes protokoly http a ftp, je nutné dodat, že prostřednictvím protokolu ftp jsou zpřístupněny také obrovské objemy dat zrcadlených ze zahraničních archivů. Proto je v případě protokolu ftp vhodné zaměřit sběr dokumentů jen na ty relevantní, tedy na dokumenty přímo odkazované ze stránek přístupných přes protokol http. V případě již zmiňovaných diskusních skupin je možné vzít v úvahu fakt, že archivy mnoha z nich jsou zároveň přístupné ve formě html-archivů dostupných také protokolem http. Pokud by se přesto ukázalo, že je důležité vytvářet jejich samostatný archiv, nabízí se k tomu standardní prostředek – instalace news serveru, který bude zrcadlit české diskusní skupiny a bude si udržovat celou jejich historii.

Podobně jako v případě protokolů bychom mohli hodnotit jednotlivé dokumenty i co do použitého formátu. Výzkumy ve světě (potvrzené i během naší dosavadní archivace českého webu) ukazují, že cca 97% počtu všech archivovaných souborů tvoří trojice formátů html, jpg a gif, ačkoli co do velikosti zaujímají soubory v těchto formátech jen asi polovinu celkového objemu dostupných dat. Pokud tedy dokážeme odpovědně určit, které ze vzácněji se vyskytujících formátů nemá smysl z různých důvodů archivovat, můžeme snadno ušetřit významnou kapacitu ukládacího prostoru, což může představovat úsporu značných částek i do budoucna. Nesmíme totiž zapomínat na to, že nestačí informace jen jednou sklidit a uložit do archivu; pro dlouhodobé zachování dostupnosti informačního obsahu každého archivovaného digitálního dokumentu (po dobu desetiletí až staletí) bude nutné zajišťovat jeho průběžnou konverzi do nových formátů, což je vzhledem k celkovému objemu dat technicky i finančně velmi náročné.

2.2 Český web

Jak již bylo uvedeno, předmětem zájmu projektu WebArchiv je archivace online publikované části

české produkce, tedy český web. V ideálním případě by výsledkem projektu měl být archiv obsahující pokud možno vše, co kdy bylo v rámci českého webu publikováno. Proto se provádí archivace dvěma cestami: *plošnou archivací*, kdy se s delším časovým odstupem (například 2krát ročně) vytváří co nejúplnější snímky celého českého webu, a *výběrovou archivací*, kdy se naopak velmi často (v případě potřeby i každý den) doplňuje archiv zrcadlící vybranou omezenou skupinu nejvýznamnějších českých zdrojů.

Aby bylo možné oba postupy realizovat, je nutné nejprve stanovit, jaký je vlastně *rozsah českého webu*. Ačkoli jej můžeme zjednodušeně definovat jako „všechny dokumenty publikované v doméně .cz," je zřejmé, že toto kritérium nepokrývá celou českou online produkci. Je vhodné rozšířit tento rozsah o mnoho dalších kategorií: dokumenty v doménách druhé úrovně registrovaných subjektem sídlícím v České republice; dokumenty publikované na serverech fyzicky umístěných v ČR; dokumenty v českém jazyce; dokumenty českých autorů; dokumenty se vztahem k Česku, atd.

V doméně .cz je nyní registrovano téměř 135.000 domén 2. úrovně. Přidáváním dalších podmínek stoupá jak náročnost nalezení všech dokumentů podmínky splňujících, tak i náročnost prokázání, že nalezený dokument některou podmínku opravdu splňuje.

2.3 Výběrová archivace

Jakmile jsme si stanovili (alespoň přibližně) rozsah českého webu, můžeme v jeho rámci začít hledat podmnožinu zdrojů, kterou by bylo vhodné archivovat výběrově – s co nejkratší periodicitou a v co největší úplnosti. V současné době se nabízí několik způsobů, jak tuto činnost zajišťovat; nejperspektivnějším z nich by mohlo být využití potenciálu projektu Jednotné informační brány CASLIN (`www.jib.cz`). Jedním z jejích výstupů bude totiž průběžně aktualizovaný předmětově členěný informační portál online elektronických zdrojů. Správa jednotlivých oborů tohoto portálu bude svěřena vždy té knihovně, která má v daném oboru největší zkušenosti. Díky tomu lze očekávat, že každý obor bude v portálu reprezentován i nejvýznamnějšími národními informačními zdroji, které se tak stanou i předmětem zájmu projektu WebArchiv.

Je zřejmé, že takto pojatý systém může mnoho serverů neoprávněně vyloučit, na druhou stranu je nutno mít na zřeteli to, že každý zdroj, zahrnutý do skupiny pro intenzivní výběrové sklízení s sebou nese nemalý díl kvalifikované lidské práce spojené s jeho knihovnickým popisem, který může ve vybraných případech jít až na úroveň jednotlivých dokumentů. Finanční náročnost může být v takovém případě samozřejmě snížena, dojde-li k nějaké formě dohody o spolupráci s příslušným vydavatelem.

2.4 Plošná archivace

Plošná automatizovaná sklizeň se snaží o co nejúplnější pokrytí národního webu v podobě časových snímků (snapshots) jednou či několikrát za rok. Volbou nejvhodnějšího nástroje pro plošnou archivaci webu se v současné době zabývá několik projektů v různých evropských zemích; za všechny zmiňme alespoň testovací projekty v Rakousku a v Dánsku (`www.netarkivet.dk`). Námi používaný produkt *NEDLIB Harvester*, vyvinutý Helsinskou národní knihovnou, ve srovnávacích testech rozhodně nezaostává. Díky tomu, že byl navržen pro potřeby archivace webu národními knihovnami, vyhovuje dobře i našim požadavkům. Nabízí velkou škálu různých nastavení, mezi něž patří volba seznamu výchozích webových stránek, omezení rozsahu sklizně pomocí URL nebo jejich částí, povolení nebo zakázání podpory protokolu ftp, logování zamítnutých URL, akceptování omezení pro roboty na jednotlivých serverech (robots.txt), podpora sklízení URL s parametrem, stanovení maximální hloubky zanoření hypertextových odkazů v rámci jednoho serveru a další. Zvláště poslední dva parametry mohou velmi významně ovlivnit rozsah a kvalitu sklizně.

Podpora URL s parametry umožňuje omezit sklízení jen na ta URL, která neobsahují znak ? uvozující seznam parametrů. Díky tomu lze sice do značné míry zabránit problémům spojeným s nekonečnými smyčkami při procházení serverů, na

druhou stranu se tak nepříjemně omezuje rozsah sklizně. Jako typický příklad lze uvést server root.cz, jehož jedinou stránkou, na kterou se dá dostat pomocí URL bez parametru, je jeho hlavní stránka. Protože podobně funguje většina elektronických periodik, vyřadili bychom ignorováním URL s parametry právě ty zdroje, které jsou z hlediska našeho kulturního dědictví nejcennější.

Je samozřejmě pravděpodobné, že mnohé dynamicky generované stránky se v archivu vyskytnou několikrát jen proto, že se navzájem nepatrně liší. Může se tak stát, že se opakovaně archivují již navštívené stránky jen proto, že součástí URL je například identifikátor sezení, nebo aktuální čas. Takový cyklus se pak opakuje tak dlouho, dokud není vyčerpán povolený počet zanoření v rámci jednoho serveru (nyní se operuje s hodnotou 50, která by měla zajistit stažení všech stránek z většiny serverů). Je však nutno poznamenat, že k podobným problémům dochází pouze v případě, kdy správce daného serveru ve vlastním zájmu nezakáže v souboru robots.txt všem robotům přístup na problematická URL.

Je zřejmé, že ať už je pro archivaci webu zvolen jakýkoli produkt, bude jím vytvořený archiv poplatný jeho limitům. Ani NEDLIB Harvester není v tomto směru samozřejmě výjimkou a tak existuje několik prozatím nepřekročitelných omezení. Jeho nejbolestivějším omezením je absence podpory javascriptu. V důsledku toho v archivu zcela chybí stránky, na něž vedou jen odkazy generované javascriptem až v prohlížeči (typickým příkladem takových odkazů jsou odkazy do archivu Neviditelného psa). Zatím méně palčivým nedostatkem stejného charakteru je absence podpory odkazů z prezentací ve formátu flash.

3 Dlouhodobé uchování a zpřístupnění zdrojů

Problematika archivace webu zahrnuje tři oblasti: první z nich je problematika automatizovaného (plošného či výběrového) sklízení informací nacházejících se na definovaném výseku webu a jejich uložení do archivu. Druhou je problematika provozování archivu, včetně konverzí formátů v něm uložených dokumentů při každé větší technologické změně. Třetí oblast pak představuje zpřístupnění informací uložených v takto vytvořených (a objemem dat velmi rozsáhlých) archivech.

3.1 Sklizeň českého webu

V loňském roce probíhala po několik měsíců v pořadí již druhá testovací sklizeň domény .cz, která bude po přestávce spojené s přechodem na nový server v letošním roce pokračovat. Tato sklizeň by měla ukázat mimo jiné i to, jaký je skutečný rozsah viditelného českého webu. Výchozími body pro tuto sklizeň byly především hlavní stránky internetových portálů seznam.cz a quick.cz. Přes různé problémy se již podařilo stáhnout z 10.490.000 URL celkem 10.090.000 souborů o souhrnné velikosti přes 240 GB. Alespoň jednou přitom bylo navštíveno přes 30.000 domén 2. úrovně (tj. čtvrtina domén v doméně .cz).

Analýza dosavadního průběhu sklizně ukazuje, jaké informační bohatství český web vlastně skrývá. Mezi padesáti našimi objemem nebo počtem souborů největšími doménami druhé úrovně najdeme mimo jiné šest univerzit, jeden univerzitou provozovaný specializovaný server (linux.cz), Českou akademii věd a několik zpravodajských a vydavatelských serverů. Dále jsou pak na předních místech zastoupeny především webhostingové farmy, které sice přináší jen minimum vlastního obsahu, ale o to větší rozmanitost.

3.2 Provoz archivu

Velikost Harvesterem tvořeného archivu může snadno dosáhnout obrovských rozměrů: jedno kolo stahování představuje v našich podmínkách stovky GB. Archiv s tak velkým potenciálem růstu není samozřejmě snadné ani levné provozovat. Ačkoli v současné době již jsou na trhu levné pevné disky o kapacitách okolo 200 GB, infrastruktura archivu se musí opírat o robustní a dlouhodobě perspektivní řešení. Toto řešení musí brát v potaz nejen aspekty technické, ale i finanční a personální a musí být z provozního hlediska dlouhodobě provozovatelné.

V pilotní fázi projektu bylo s výhodou využito stávajícího páskového robota Národní knihovny

ČR; jeho nevýhodou ovšem je problematická dostupnost na něm uložených dat v okamžiku, kdy by bylo nutné tato data zpřístupnit veřejnosti. Protože stažené dokumenty jsou společně s příslušnými metadaty ukládány jako tar+gzip komprimované soubory přímo do souborového systému, neměl by být problém s migrací dat na nová úložiště.

Větším oříškem samozřejmě bude zajištění technologické čitelnosti archivovaných souborů. Je sice pravděpodobné, že nejrozšířenější otevřené formáty (html, txt, gif, jpg) zůstanou interpretovatelné po velmi dlouhou dobu, oprávněné pochybnosti lze však mít o dlouhodobé čitelnosti proprietárních formátů – především těch z nich, které nejsou tak rozšířeny jako například formáty firem Adobe nebo Microsoft. I u formátů Microsoftu je však zárukou jejich budoucí interpretovatelnosti spíše dostupnost alternativních programů s otevřeným kódem, které umí s těmito formáty pracovat (OpenOffice), než vlastní podpora ze strany Microsoftu.

Ať už bude v budoucnosti vývoj tohoto archivu jakýkoli, lze říci, že využitím NEDLIB Harvesteru získala Národní knihovna vhodný nástroj pro tvorbu konzervačního archivu českého webu. Vytvoření a udržování takového archivu je důležitým krokem na cestě k naplnění jeho smyslu, tedy ke zpřístupnění obsahu archivu uživatelům.

3.3 Zpřístupnění informací v archivu

Pro zpřístupnění dokumentů v archivu se nabízejí technologie fulltextového indexování a automatizované extrakce autorem vytvořených metadat. Na naši zakázku byl koncem roku 2001 vypsán na MFF UK ročníkový týmový projekt na vytvoření *indexační a vyhledávací aplikace pro WebArchiv*. Tato aplikace by měla zpřístupnit stažené dokumenty v jejich kontextu, tedy s vloženou grafikou ze stejné doby a s odkazy vedoucími primárně opět do archivu. Vyhledávání v archivu by mělo být možné nejen na základě zadání předem známého URL nebo kontrolního součtu dokumentu, ale i na základě metadat extrahovaných z dokumentu nebo fulltextového vyhledávání. Aplikace bude navržena tak, aby bylo možné připojit k ní moduly pro indexování i jiných než textových dokumentů – jeden z takových nástrojů, Convera Retrievalware, je již v NK zkušebně provozován. Jedním z budoucích cílů projektu bude proto pokus o jeho využití pro indexování některých netradičních typů souborů obsažených v archivu.

Nadějně se jeví též kontakty s týmem Norské národní knihovny, který vyvinul a v letošním roce se chystá dát volně k dispozici vlastní systém pro indexaci a zpřístupnění webového archivu založený na indexovacím stroji Apache Jakarta Lucene.

4 Perspektiva projektu

To, zda bude některá z dosud popisovaných ověřených technologií nasazena také v ostrém reálném provozu v rámci České republiky, bude záviset i na vyřešení autorskoprávní problematiky související s tvorbou a provozem webového archivu. Nedotaženost zákona o povinném výtisku u nás otevírá cestu různým výkladům omezení daných zákonem o autorském právu. Automatickou identifikaci a archivaci online publikovaných dokumentů lze srovnat s dnes běžně používanou technologií indexování webu, jak ji provádějí Internetové vyhledavače. Bez konkrétní opory v zákoně ale není jisté, zda bude možné stávající strategii plošné archivace využívat i pro zpřístupnění shromážděných dat. Existující infrastruktura je nastavitelná tak, aby bylo možné zachovat alespoň omezený rozsah sklízení i v případě, že bude nutné podřídit se určitým zákonným omezením. Závažným důsledkem takových omezení by však bylo velmi výrazné zmenšení rozsahu sbírky, tvořené pak víceméně jen na základě dobrovolně dodávaných dokumentů. Je možné prohlásit, že právo občana na informace by mělo být naplněno i existencí digitální knihovny obsahující elektronicky publikované dokumenty v nezměněné podobě.

Ačkoli je díky vytvořené infrastruktuře již nyní možné udělat mnohé pro zachování soudobých informačních zdrojů pro budoucí generace, další rozvoj této infrastruktury, stejně jako vývoj všech podpůrných softwarových produktů, nemůže být nikdy zcela ukončen. Zde nejde jen o hledisko potřeb uživatele nebo provozovatele,

ale i o hledisko technického vývoje, mezinárodní spolupráce nebo problematiku legislativní. S tím, jak bude stoupat podíl čistě elektronické produkce, bude růst i význam její dlouhodobé archivace z hlediska ochrany národního kulturního dědictví. I proto je žádoucí, aby projekt WebArchiv, i přes nevyjasněnou legislativní situaci, ve své činnosti pokračovat. □