Projekt WebArchiv – archiv českého webu

Adam Brokeš, ÚVT a FI MU

Webové médium patří mezi ta nejdynamičtěji se vyvíjející, a také ta nejkřehčí. Podle některých studií je životnost elektronických dokumentů na webu necelých 100 dní. S přihlédnutím k tomu, že až 90% těchto dokumentů existuje pouze v digitální podobě, není těžké si představit budoucnost, ve které naši potomci budou hledět na dnešní období jako na dobu „digitálního temna". Z těchto důvodů po celém světě vznikají instituce, které se zabývají archivací a zpřístupňováním (nejen) webových dokumentů. Protože tato činnost je analogická klasickým knihovnickým službám, vytvořila se specializovaná oddělení zabývající se danou problematikou především v rámci řady národních knihoven. Ze soukromých hráčů na tomto poli jmenujme Internet Archive `www.archive.org` – tento archiv shromažďuje data již od roku 1996 a velikost se pohybuje v řádech desítek petabytů. Všechny tyto instituce spojuje konsorcium IIPC – International Internet Preservation Consortium, které koordinuje spolupráci mezi jednotlivými členy. Národní knihovna České republiky je členskou organizací od počátku roku 2007.

1 Projekt WebArchiv

Úlohou projektu WebArchiv `http://www.webarchiv.cz` je řešení problematiky archivace národního webu, tj. bohemikálních dokumentů zveřejněných v prostředí sítě Internet. Jde o shromažďování webových zdrojů, jejich archivaci, ochranu a zajištění dlouhodobého přístupu. Provádí se jednak kompletní plošná archivace, tj. automatický sběr „celého" českého webu, souběžně však probíhá i výběrová archivace (nejzajímavějších webových zdrojů vybraných na základě selekčních kritérií) a tematické archivace (zaměřené na určité aktuální téma, např. volby, povodně apod.). V současné době je stav řešení na úrovni funkčního provozu s testováním nových funkcí. K převedení do plně rutinních činností je zapotřebí jednak podstatné navýšení financování projektu, jednak změny stávající legislativy (zejména autorsko-právní) tak, aby umožňovala zpřístupňování archivovaných zdrojů.

2 Historie

WebArchiv vznikl v rámci programového projektu výzkumu a vývoje „Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet" pod záštitou Ministerstva kultury ČR. Projekt je řešen od roku 2000 v Národní knihovně České republiky a financován téměř výhradně z grantové podpory. Spoluřešitelem odpovědným za informační technologie je Moravská zemská knihovna v Brně (MZK), externím spolupracovníkem je Ústav výpočetní techniky Masarykovy univerzity v Brně (ÚVT). Na programovém řešení se podílí tým studentů Fakulty informatiky MU. V roce 2000 byl projekt technicky zajištěn jedním serverem umístěným v MZK a páskovým robotem, který se nacházel v Národní knihovně. Sklízení probíhalo nástrojem NEDLIB Harvester, robotem vyvíjeným Helsinskou národní knihovnou. Tento robot sloužil dobře pro výběrové sklizně, ale při celoplošném sklizení domény .cz jsme narazili na technické obtíže. Robot se po čase zpomalil do té míry, že nebylo možné dále pokračovat ve sklizni. Dnes je již vývoj zastaven. V roce 2004 byl nahrazen Heritrixem, open-source crawlerem vyvíjeným pod záštitou Internet Archive, a výkonnějšími servery.

3 Současný stav

3.1 Workflow

V současné době je workflow rozdělena na technickou a logickou část. Pracovníci v Národní knihovně zajišťují výběr a hodnocení zdrojů, jejich katalogizaci a kontaktování vydavatelů. Dále vytvářejí popisná metadata (Dublin Core), jsou důležitým spojovacím článkem mezi vydavateli a technickou podporou v Brně a vytvářejí podklady pro prezentaci projektu, především obsah pro webové stránky. V Praze je také umístěn server zpřístupňující archív a webový portál projektu `http://www.webarchiv.cz`.

Brněnská část týmu se stará o technické zázemí projektu. Jsou zde umístěny dva servery. Probíhá zde sklízení dat, provoz interního systému, vývoj

a testování. Zároveň je třeba udržovat hardware, jeho provoz a provádět údržbu a lokalizaci použitého software.

3.2 Provedené sklizně, popis archivu

Celoplošné sklizně. Sklizeň probíhá na celé doméně „.cz", dnes je seznam domén druhé úrovně získáván od registrátora NIC.cz. Úkolem sklizně je zachytit co nejširší rozsah bohemikálních dokumentů.

- 2001 – První pokus o provedení celoplošné sklizně pomocí jednoho serveru s páskovým robotem, sklizeň nedokončena díky technickým problémům. Hloubka zanoření 25 odkazů.
- 2002 – Sklizeň byla přerušena z důvodu záplav a fyzického zatopení serveru umístěného v Národní knihovně.
- 2004 – Sklizeň proběhla úspěšně, zastavena byla po zaplnění diskového prostoru. Hloubka zanoření 50 odkazů.
- 2005 – První sklizeň provedena pomocí robota Heritrix. Zastavena po havárii robota, která byla způsobena nedostatky tehdejší verze.
- 2006 – Sklizeň pomocí Heritrixu, pozastavena po zaplnění diskového prostoru. Byl nastaven limit 100 MB na soubor a 5 000 dokumentů na server.
- 2007 – Zatím nejúspěšnější sklizeň, bylo sklizeno 81,3 mil. dokumentů o celkové komprimované velikosti 3,6 TB. Vstupem bylo 320 tisíc domén druhé úrovně a celý proces trval necelý měsíc.

Výběrové sklizně. Tyto sklizně probíhají periodicky několikrát ročně na základě výběru určitého zdroje, který splňuje selekční kritéria. Výběr probíhá v Národní knihovně a posléze je kontaktován vydavatel zdroje, který, pokud souhlasí, podepíše smlouvu. Sklizený materiál, který již je v archivu umístěn nebo bude do něj zařazen v budoucnosti, je možné legálně zpřístupnit. Těchto smluv je v současné době přes 460.

Tematické sklizně. Při tomto druhu sklizně je zacílena množina stránek týkajících se zvoleného tématu. Dosud proběhly sklizně: Dalimilova kronika, Povodně 2002, Vysočina, Volby 2006, Prezidentské volby 2008, Nová budova NK, Nová budova Národní technické knihovny, Praha olympijská.

Statistika archivu. V současné době je v archivu uloženo 8,9 TB nekomprimovaných dat, což činí přibližně 200 milionů dokumentu. Celých 70 % je tvořeno HTML soubory, které se dají velice efektivně komprimovat.

3.3 Nástroje

Heritrix. Heritrix je open-source sklízecí robot (crawler), který je vyvíjen společností Internet Archive. Je velice modulární, rozšiřitelný a nezávislý na platformě (je napsán v jazyce Java). Skládá se z frameworku (samotného jádra programu) a modulů (frontiers, processors, scopes, filters). Samotné nastavení heritrixu je vytvoření konkrétního zapojení a zřetězení modulů. Tímto řetězcem poté projde každý URI (Uniform Resource Identifier) a je zpracován podle zapojených modulů. Musím ocenit kvalitní a rychlou pomoc ze strany vývojářů heritrixu a podrobnou dokumentaci. V současné době je k dispozici verze 1.12.1, která se zaměřila zkvalitnění ochrany před pádem do pasti (dynamicky generované stránky na kterých se může robot zacyklit) a vnitřní deduplikaci.

Koncem února vyšla v druhé vývojové větvi verze 2.0.0. Ačkoliv je to již finální verze, není v projektu v současnosti využívána, protože skok mezi verzemi je značný a především není zajištěna kompatibilita mezi balíky nastavení pro jednotlivé sklizně a tak bychom nemohli automatizovanou formou vyžít zkušeností z již funkčních sklizní. Nová verze ale přináší velké množství užitečných novinek a je tedy jisté, že po vyřešení počátečních problému bude nasazena. Velkou změnou je striktní oddělení webového rozhraní a samostatného sklízecího robota (komunikace probíhá pomocí JMX technologie). Takto není problém ovládat několik robotů z jedné adresy. Přepracován byl i systém definice nastavení pro jednotlivé domény (pokud například nechceme sklízet část domény) a možnost prioritizace front.

DeDuplicator. Je modul, který umožňuje vytvoření indexu sklizených dat (z předchozích logů

nebo během sklizně) a při dalším sklízení porovnává data zařazená ve frontě s těmi, co se již nacházejí v indexu. Lze tak zamezit ukládání duplicitních dat a dokonce je možné tato data vyřadit z fronty ještě před jejich stažením. Využívá se především pro méně často se měnící dokumenty binárního charakteru (obrázky, video, zvuk). Formát ARC, do kterého ukládá data Heritrix, neumožňuje plně využívat možnosti DeDuplicatoru, např. možnost odkazovat na dokument stažený z jiného URL. Tyto nedostatky by měl odstranit nástupce ARCu – WARC.

WARC. Formát ARC, který je nyní používán pro archivaci dat, se ukázal nedostačující možnostem, které nabízí harvester Heritrix a požadavkům kladeným knihovnickým personálem. Z těchto důvodů vznikl nový formát WARC (Web ARChive). Data jsou podobně jako v předchozím formátu ukládána spolu s hlavičkou sekvenčně za sebou, spojována do tar balíků a ty jsou posléze komprimovány GZ kompresí. WARC ale značně rozšiřuje hlavičku, nyní je tedy možné uložit kompletní informace od Heritrixu (request i reply protokolu atd.), označovat duplikované soubory, odkazovat mezi již uloženými soubory atd. Hlavičku lze podle XML Schematu dále rozšiřovat o libovolná metadata. Podpora v Heritrixu je finální až ve verzi 2, tedy přechod na tento formát se uskuteční spolu s využitím nové verze.

Wayback. Open-source aplikace vyvíjená v jazyce Java společností Internet Archive pro zpřístupnění archivovaných dokumentů koncovým uživatelům, která nahradila původní Wayback Machine použitý přímo na stránkách archive.org. Dokumenty jsou indexovány a zpřístupňovány pomocí URL. Je implementována podpora pro hvězdičkovou notaci. Systém může pracovat ve třech módech:

- Archival URL – systém pomocí javascriptu změní url odkazy na stránce tak, že odkazují zpět do archivu.
- Proxy – systém se chová jako proxy server, je obtížné měnit časové verze.
- Timeline – u serveru vždy zobrazí časovou osu, tato funkce je experimentální.

V přípravě je fulltextové vyhledávání a lokalizace. V tuto chvíli je wayback využit ve WebArchivu pro zpřístupnění celého archivu – avšak pro zobrazení obsahu, na který není smlouva, je třeba přistupovat z počítačů Národní knihovny (volný přístup odkudkoliv současná česká legislativa neumožňuje).

NutchWAX. Tato nadstavba vyhledávacího systému Nutch byla vytvořena přímo pro potřeby indexování dokumentů archivovaných Heritrixem (ARC formátu). Přidává do formátu potřebná metadata, především časové razítko. V této chvíli je vydána verze 0.10, která podporuje zpracování velkých objemů dat a distribuovaný file systém Hadoop. V projektu WebArchiv je použit pro indexování smluvních zdrojů.

Netarchive Suite. Systém vyvíjený původně pro potřeby dánské národní knihovny. Umožňuje automatizovanou sklizeň definovaných kolekcí URL a kontrolu kvality. Systém je dekomponován na množství nezávislých modulů, které komunikují pomoci JMS technologie. Sklízení pomocí tohoto software by bylo možné i technicky neerudovaným personálem. Chybějící podpora nasmlouvaných zdrojů (pro které je získána smlouva od vydavatele) nám ale stále brání v nasazení. Jednou z možností je úprava aplikace pro naše účely (zdrojový kód je vydán jako open-source), případné provázání s novou verzí WA Adminu, kterou knihovníci pro tuto správu využívají a já sám ji řeším v rámci své bakalářské práce.

4 Závěr a zhodnocení

Projekt typu WebArchiv je obtížné hodnotit v současnosti, jeho hlavní přínos se projeví až ve vzdálenější budoucnosti. Z dnešního pohledu je důležité, že Česká republika se zařadila mezi nejvyspělejší země světa, které jsou ochotny a schopny danou problematiku řešit, jakkoliv je to dnes ještě poměrně obtížné – ať již z pohledu technologického, legislativního či finančního.

Pro mne je na projektu největším přínosem právě uvědomění si velmi krátkého poločasu rozpadu elektronických informací a hledání cest jak tomuto rozpadu předejít. Věřím, že zanecháváme poselství a užitnou hodnotu budoucím generacím, které jednou nebudou muset hledět na naši dobu jako na informační černou díru.

Literatura

[1] http://www.webarchiv.cz/
[2] http://www.archive.org/
[3] http://netarchive.dk/
[4] http://crawler.archive.org/
[5] http://archive-access.sourceforge.net/projects/wayback/
[6] http://deduplicator.sourceforge.net/

□