WebArchiv - archiv českého webu

http://www.webarchiv.cz

Adam Brokeš, FI MU -2. ročník

1. Popis projektu

Úlohou projektu WebArchiv je řešení problematiky archivace národního webu, tj. bohemikálních dokumentů zveřejněných v prostředí sítě Internet – shromažďování webových zdrojů, jejich archivace a ochrana a zajištění dlouhodobého přístupu k těmto archivovaným dokumentům. Provádí se jednak kompletní archivace, tj. automatický sběr „celého“ českého webu. Souběžně probíhá výběrová archivace (na základě URL nejzajímavějších webových zdrojů vybraných na základě selekčních kritérií) a tématické archivace (zaměřené na určité aktuální téma, např. volby, povodně apod.). V současné době je stav řešení na úrovni výzkumu a testování. K provádění rutinních činností je zapotřebí jednak podstatné navýšení financování projektu, jednak řešení stávající legislativy zejména autorsko-právní tak, aby umožňovala zpřístupňování archivovaných zdrojů. Protože jsem součástí tohoto projektu, budu čerpat i z vlastních 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, 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). 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 ve sklizni pokračovat. Dnes je již vývoj zastaven. V roce 2004 byl nahrazen Heritrixem, open-source crawlerem vyvíjeným pod záštitou Internet Archive. Postupně přibyly další dva servery, umístěné v ÚVT a páskový robot byl nahrazen diskovým polem o kapacitě 4,6 TB.

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.

Brněnská část týmu se stará o technické zázemí projektu. Jsou zde umístěny tři servery. Probíhá zde sklízení dat, zpřístupňování sklizených dat, úpravy webu, vývoj a testování. Zároveň je třeba udržovat provoz hardware a provádět údržbu a lokalizaci použitého software.

3.2 Provedené sklizně, popis archivu

3.2.1 Celoplošné sklizně

Sklizeň probíhá na celé doméně „.cz“, dnes je seznam domén druhé úrovně získáván za poplatek od registru NIC.cz. Úkolem sklizně je zachytit, co nejširší rozsah bohemikálních dokumentů.

- 2001 – První pokus o provedení celoplošné sklizně pomocí jednoho serveru s páskovým robotem, nedokončena díky technickým problémům. Hloubka zanoření 25 odkazů
- 2002 – Sklizeň byla přerušena z důvodu záplav a fyzického zatopení serveru umístěného v Národní knihovně.
- 2004 – Zastavena 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 – Úspěšná sklizeň pomocí Heritrixu, pozastavena po zaplnění diskového prostoru. Byl nastaven limit 100MB na soubor a 5000 dokumentů na server

3.2.2 Výběrové sklizně

Tyto sklizně probíhají na základě výběru určitého zdroje, který splňuje selekční kritéria. Tento výběr probíhá v Národní knihovně a posléze je kontaktován vydavatel zdroje, který, pokud souhlasí, podepíše smlouvu a materiál, který je již umístěn, nebo bude umístěn v budoucnosti do archivu je možné legálně zpřístupnit. Těchto smluv je v současné době přes 300.

3.2.3 Tematické sklizně

Při tomto druhu sklizně je zacílena určitá množina stránek týkajících se jednoho tématu. Dosud proběhly sklizně: Dalimilova kronika, Povodně 2002, Vysočina, Volby 2006.

3.2.4 Statistika archivu

V současné době je v archivu uloženo 5,6TB nekomprimovaných dat, což činí přibližně 135 milionů dokumentu. Celých 70% je tvořeno HTML soubory, které se dají velice efektivně komprimovat.

3.3 Nástroje

3.3.1 Heritrix

Heritrix je open-source sklízecí robot (crawler), který je vyvíjen společností Internet Archive. Je velice modulární, rozšiřitelný a platformě nezávislý (je napsán v jazyce Java). Skládá se z frameworku (samotného jádra programu) a modulů (frontiers, processors, scopes, filters). Samotné nastavení heritrixu je vlastně vytvoření konkrétního zapojení a zřetězení modulů.

Tímto řetězcem poté projde každé URI a je zpracováno podle zapojených modulů. Musím ocenit kvalitní a rychlou pomoc ze strany vývojářů heritrixu a podrobnou dokumentaci. V současné době je verze 1.10.1, která se zaměřila na zvýšení bezpečnosti samotného interface a zkvalitnění ochrany před pádem do pasti (dynamicky generované stránky na kterých se může robot zacyklit). Bohužel i nadále není možné provádět celoplošnou sklizeň bez odborných zásahů během sklizně.

3.3.2 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 tyto 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.

3.3.3 WERA

Wera vznikla za spolupráce konsorcia IIPC, Internet Archive a NWA jako nástroj sloužící pro zpřístupnění archivu. Byl vyvíjen v PHP a má velice propracované uživatelské rozhraní. Obsahuje například časovou osu, na které si lze vizuálně zobrazit časové verze dokumentu. Archivovaný dokument lze zobrazit i zadaním přesného URL. WERA využívá index NutchWax a lze tak v ní fulltextově vyhledávat. Občas se vyskytne chyba při spouštění Javascriptu. Dnes je již vývoj ukončen. V projektu WebArchiv je WERA využita pro zpřístupnění výběrových sklizní, na které je podepsána smlouva. I tento stav brzy pomine a celé zpřístupnění se přesune na Wayback.

3.3.4 Wayback

Tato open-source aplikace vyvíjená v jazyce Java společností Internet Archive má v blízké budoucnosti nahradit 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 podobně jako ve WERA, tato funkce je experimentální.

V přípravě je fulltextové vyhledávání a lokalizace. V tuto chíli je wayback využit ve WebArchivu pro zpřístupnění celého archivu, avšak pro zobrazení obsahu, na který není smlouva je třeba na server přistupovat z Národní knihovny.

3.3.5 Nutch

Je open-source vyhledávací engine, který vznikl jako reakce na uzavřené komerční vyhledávací systémy. Umí stáhnout miliony stránek měsíčne, spravovat jejich obsah a vyhledávat v něm
1000x za vteřinu.

3.3.6 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 nestabilní verze 0.6, která podporuje zpracování velkých objemů dat a distribuovaný file systém Hadoop. V projektu Webarchiv je použit pro indexování smluvních zdrojů.

3.3.7 Webcurator

První verze tohoto open-source softwaru byla představena na konferenci IWAW v září 2006. Systém vznikl díky spolupráci Britské knihovny a Národní knihovny Nového Zélandu. Vznikl z jednoduché potřeby. Umožnit i netechnickému personálu knihovny provádět sklizně vybraných zdrojů s minimálními technickými znalostmi prostřednictvím přívětivého a propracovaného grafického prostředí. Bohužel systém v tuto chvíli nepodporuje inkrementální sklízení a konfigurace nástroje není zcela konzistentní.

4. Závěr a zhodnocení

Projekt typu WebArchiv nelze hodnotit ihned po jeho vytvoření. Přínos se odrazí až ve vzdálenější budoucnosti, tedy za čas, který v IT není úplně obvyklý. Pro mne je na projektu největším přínosem právě uvědomění si velmi krátkého poločasu rozpadu informací a hledání cest jak tomuto rozpadu předejít. Věřím, že zanecháváme poselství a užitnou hodnotu budoucím generacím, které jednou nebudou muset hledět na minulost jako na černou díru.

Zdroje, literatura

http://www.webarchiv.cz/
http://www.archive.org/
http://webcurator.sourceforge.net/
http://crawler.archive.org/
http://archive-access.sourceforge.net/projects/wayback/
http://deduplicator.sourceforge.net/

Metadata v Dublin Core

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="WebArchiv" />
<meta name="DC.Creator" content="Adam Brokeš" />
<meta name="DC.Subject" content="WebArchiv" />
<meta name="DC.Subject" content="archivace webu" />
<meta name="DC.Description" content="Úlohou projektu WebArchiv je řešení problematiky archivace národního webu, tj. bohemikálních dokumentů zveřejněných v prostředí sítě Internet ." />
<meta name="DC.Date" content="2.1.2006" />
<meta name="DC.Type" content="Text" />
<meta name="DC.Type" content="esej" />
<meta name="DC.Format" content="application/doc" />

<meta name="DC.Format" content="computerFile" />
<meta name="DC.Identifier" content="http://www.fi.muni.cz/~xbrok/files/webarchiv.doc" />
<meta name="DC.Source" content="http://www.webarchiv.cz/" />
<meta name="DC.Source" content="http://www.archive.org/" />
<meta name="DC.Source" content="http://webcurator.sourceforge.net/" />
<meta name="DC.Source" content="http://crawler.archive.org/" />
<meta name="DC.Source" content="http://archive-access.sourceforge.net/projects/wayback/" />
<meta name="DC.Source" content="http://deduplicator.sourceforge.net/" />
<meta name="DC.Language" content="cze" />