Masarykova univerzita
Fakulta informatiky

Nástroje pro migraci webového archivu

Bakalářská práce

Martin Prokop

Brno, jaro 2012

Prohlášení

Prohlašuji, že tato bakalářská práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Vedoucí práce: Mgr. Václav Rosecký

Poděkování

Na tomto místě chci poděkovat Mgr. Václavovi Roseckému a Ing. Petrovi Žabičkovi za pomoc při řešení této bakalářské práce.

Shrnutí

Bakalářská práce se zabývá možností migrace webového archivu vytvořeného v rámci projektu WebArchiv. Představuji projekt a jeho vývoj, dále podávám přehled nástrojů a prací s ním souvisejících.

Po přehledu a sumarizaci informací o WebArchivu pokračuji přehledem archivačních formátů. Předvádím výhody jednotlivých formátů a ukazuji důvody, proč je potřeba provést migraci stávajícího archivu.

V poslední části práce se zabývám samotnými nástroji sloužícími k migraci. Cílem práce je představit výhody a nevýhody jednotlivých nástrojů a tím připravit podklad pro další studii možnosti migrace a její realizaci.

Abstract

Bakalářská práce se zabývá možností migrace webového archivu vytvořeného v rámci projektu WebArchiv. Představuji projekt a jeho vývoj, dále podávám přehled nástrojů a prací s ním souvisejících.

Po přehledu a sumarizaci informací o WebArchivu pokračuji přehledem archivačních formátů. Předvádím výhody jednotlivých formátů a ukazuji důvody, proč je potřeba provést migraci stávajícího archivu.

V poslední části práce se zabývám samotnými nástroji sloužícími k migraci. Cílem práce je představit výhody a nevýhody jednotlivých nástrojů a tím připravit podklad pro další studii možnosti migrace a její realizaci.

Klíčová slova

WebArchiv, warc, arc, migrace, archivace webu.

Obsah

Kapitola 1

Úvod

V dnešní době je internet jedním z hlavních zdrojů a nositelů informací. Každým dnem vzniká velké množství nových elektronických dokumentů. Takové dokumenty nemají charakter stálých informací, jejich obsah se mění prakticky neustále. Je typické, že staré verze dokumentů jejich autoři neuchovávají, a proto dochází ke ztrátě cenných informací. Z toho důvodu je důležité, aby se webové zdroje dlouhodobě uchovávaly a bylo možné je zpětně rekonstruovat. Tato práce je cílem projektu WebArchiv.

Řešitelé projektu WebArchiv se archivaci věnují již dvanáctým rokem. S postupujícím zdokonalováním technologií se rychle mění trendy a parametry, které musí sledovat a splňovat. Dá se říci, že každým dalším dnem vznikají další požadavky pro archivaci a dlouhodobé uchování webových zdrojů. Jedním ze základních problému je samotné uchovávání datového archivu. Každým rokem narůstá jeho obsah a tomu se musí přizpůsobit práce s archivovanými zdroji. Dále je potřeba zajistit, aby archivovaná data byla čitelná dlouhodobě a dalo se s nimi snadno manipulovat. V současné době se projevují limity archivování zdrojů pomocí formátu arc. Tento formát je pro archivaci sice vhodný, ale je již zastaralý a málo robustní.

Ukazuje se, že by bylo vhodné nahradit arc formátem warc. Formát warc je relativně nový a prozatím ne příliš používaný. V budoucnu se však pravděpodobně stane standardem pro uchovávání webových zdrojů. Vzniká tedy otázka možnosti přechodu WebArchivu na práci výhradně s tímto formátem. Před samotným přechodem je potřeba vyřešit spoustu problémů a všechny používané nástroje musí být na práci s novým formátem připraveny.

Předmětem této bakalářské práce je jedna z dílčích akcích vedoucích k přechodu WebArchivu na uchovávání dat ve formátu warc archivů.

Jedná se o migraci stávajícího archivu do nového formátu. Tato problematika je nová a specifická přímo pro projekt WebArchiv.

Ve své práci se pokusím nalézt nástroje, které by samotnou migraci umožnili. Představím možná řešení převodu a ukáži jejich výhody a nevýhody. Výstupem práce by pak měly být ukázky jednotlivých nástrojů určených pro migraci, nikoliv samotná migrace. Závěry práce tedy budou tvořit podklady řešitelům projektu WebArchiv, pro plánování migrace webového archivu.

Kapitola 2

WebArchiv

2.1 O WebArchivu

WebArchiv je projekt, jehož cílem je archivace Českého internetu. Jeho zřizovatelem je Národní knihovna ČR, která spolupracuje s Moravskou zemskou knihovnou a Ústavem výpočetní techniky Masarykovy univerzity. Projekt vznikl roku 2000 v rámci projektu Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet. Jeho cílem je uchovat české webové zdroje v rámci zachování českého kulturního dědictví. WebArchiv klade důraz na takové elektronické zdroje, které nejsou dostupné v tištěné podobě. Jejich archivace ve WebArchivu jeden z nejspolehlivějších způsobů jejich uchování do budoucna.

Zachování těchto dokumentů je podstatné hlavně pro zachycení a možnost hodnocení vývoje českého kulturního dědictví. Projekt si klade za cíl uchovávat dlouhodobě české webové stránky a umožnit jejich zpětné vyhledávání.

Archivace není důležitá jen pro případ ztráty dat, ale i pro předejití problému zastarání technologií. Jak víme multimediální technologie se v dnešní době velmi dynamicky vyvíjí a dokument, který byl před několika lety běžně zobrazitelný pro většinu uživatelů internetu, již dnes není podporovaný většinou počítačů. Jelikož WebArchiv zálohuje velké množství dokumentů v průběhu let musí řešit i tento aspekt archivace dat. Má tudíž výborný potenciál k zachování informační hodnoty dokumentů.

Databáze WebArchivu obsahuje :

- Digitální dokumenty volně dostupné prostřednictvím sítě internet
- Publikace odborného, uměleckého a zpravodajsko-publicistického

zaměření

- Periodika, monografie, konferenční příspěvky, výzkumné a jiné zprávy, a akademické práce

nadpis1 Digitální dokumenty [1] volně dostupné prostřednictvím sítě internet

nadpis2 Publikace odborného, uměleckého a zpravodajsko-publicistického zaměření

nadpis3 Periodika, monografie, konferenční příspěvky, výzkumné a jiné zprávy, a akademické práce

2.2 Vývoj projektu

Pokusím se předvést cíle a vývoj projektu v průběhu jeho řešení. V počátcích projektu bylo nejprve potřeba stanovit základní parametry projektu a dlouhodobé cíle. V průběhu řešení pak docházelo hlavně k formulaci nových cílů, výzkumu nových technologií a vznikla i potřeba pro adaptování se na nové trendy a technologie. [1]

WebArchiv pravidelně publikuje zprávy o své činnosti a dalších záměrech. Zprávy jsou samozřejmě vystavovány na webových stránkách projektu. Přehled vývoje v jednotlivých letech, který uvedu níže, čerpal vždy výhradně ze zpráv pro příslušný rok. [1]

2.2.1 Rok 2000

V počátcích projektu bylo důležité připravit podmínky, které by umožnili samotné zpracovávání české národní bibliografie a zajistit její dlouhodobé ukládání. Následně bylo potřeba vyřešit organizační otázky týkající se získávání nových dokumentů. Zde se jedná hlavně o legislativní rámec a politiku přijímání nových dokumentů. A poté upřesnění možností přístupu k elektronickým zdrojů v souvislosti s autorským právem.

1. dokument no

2.2.2 Rok 2001

Důležitou prací bylo mapování situace s archivováním internetových zdrojů v rámci jiných projektů a institucí, protože čerpání zkušeností od jiných řešitelů podobných projektů je výhodné a může ušetřit spoustu práce do budoucna. Začalo samotné shromažďování internetových zdrojů. Zásadní byla otázka archivace sklizených dat, řešitelé potřebovali, pro dlouhodobou archivaci a práci s daty, zvolit vhodný formát uchovávání metadat stažených souborů. Zvolili Dublin Core Metadata Element Set1, který byl lokalizován pro české zdroje. Probíhal vývoj nových nástrojů: Dublin Core Metadata Generator, Generátor URN, Kalkulátor MD5, Nedlib Harvester.

Rok 20012

Řešitelé upravili kritéria výběru nových webových zdrojů. Jak jsem zmínil výše, došlo k vývoji v oblasti legislativy. Opět vycházeli ze strategie zahraničních kolegů. Dále byla věnována pozornost sklizni elektronických seriálů.

pomlčka - spojovník – dlouhán —
italika *neco*
strojopis `neco - link a tak`
uvozovky? „noco“
sdgew

Literatura

[1] Co je WebArchiv?. WEBARCHIV. *WebArchiv* [online]. 21.3.2012 [cit. 2012-03-21].Dostupné z: http://www.webarchiv.cz.

Kapitola 3

Přílohy

Rok 20012

to co má hvězdičku totiž neni v obsahu