Software pro archivaci webu

Ing. Petr Žabička, MZK
Mgr. Lukáš Matějka, MU
Adam Brokeš, MU

1. Specifika webového archivu

- velmi velké množství souborů
- vysoké objemy dat
- velké množství datových formátů
- automatizovaná akvizice dokumentů
- neustálý vývoj standardů, protokolů,...

2. Možnosti řešení

2.1 online archiv (Internet Archive)
2.2 využití online služby
2.3 programy pro archivaci „malého“ rozsahu
2.4 sada profesionálních nástrojů

2.1 Online archiv

- archive.org, webarchiv.cz, ...
- pasivní využití existujících služeb
- omezená dostupnost dokumentů
- závislost na rozhodnutích provozovatele služby

2.2 Online služby

- hanzoweb.com, hanzoarchives.com
- zdarma do 100 MB měsíčně
- prvky Webu 2.0
- spolupráce uživatelů na tvorbě archivu
- služby pro jednotlivce i firmy

2.3 Programy pro archivaci

- httrack, wget, Teleport,...
- náročnější na používání
- Vhodné pro jednotlivce a menší projekty
- při velkých objemech dat problémy se škálovatelností

2.4 Profesionální nástroje pro rozsáhlé projekty

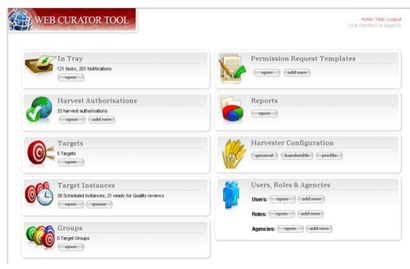
- konsorcium International Internet Preservation Consortium (IIPC - www.netpreserve.org)
- od 2007 členem i Národní knihovna ČR
- jedním z dominantních členů je Internet Archive, dále převážně velké národní knihovny
- konsorcium vyvíjí sadu open source nástrojů pro stahování, archivaci a zpřístupnění webu
- uživatelská základna členů IIPC zaručuje dlouhodobý rozvoj vyvíjených nástrojů
- IIPC nabízí mechanismy pro podporu vývoje nových aplikací podle potřeb svých členů
- Standardizace – vývoj formátu WARC pro archivaci webu

2.4.1 Heritrix

- web harvester (nástroj pro stahování webu)
- modulární, rozšiřitelný, probíhá neustálý vývoj (nyní verze 1.12.1)
- robustní (podpora plošného stahování webu)
- platformě nezávislý (java aplikace)
- snaha o co největší eliminaci stahování duplikátů a nepotřebných dat (úspora 40-60% místa oproti dřívějšku)
- obsahuje již i parser pracující s modelem dokumentu (kvalitnější analýza dokumentu, ale paměťově náročné)

2.4.2 Web Curator Tool

- nástroj pro správu sklízení, první verze uvolněna v září 2006
- nadstavba pro Heritrix
- umožňuje správu sklízení méně kvalifikovaným uživatelům prostřednictvím graficky přívětivého a propracovaného webového rozhraní
- výborná podpora uživatelských oprávnění
- nepodporuje inkrementální sklízení
- nekonzistentní konfigurace
- implementovaný workflow nepodporuje možnosti dané naší legislativou



2.4.3 Nutch

•volně dostupný modulární vyhledávací engine
• umí stáhnout a zpracovat miliony stránek měsíčně; spravovat jejich index, vyhledávat v něm 1000x za vteřinu
• snadná škálovatelnost (využití systému Hadoop pro rozložení aplikace na více serverů)

2.4.5 WERA

• WEb aRchive Access – zpřístupnění webového archivu
• velmi snadná navigace a propracované uživatelské rozhraní (časová osa zobrazuje časové verze dokumentu)
• zobrazovat archivované stránky lze i pomocí zadání přesné URL adresy
• archivované dokumenty a WERA propojeny skrz index NutchWAXe
• problémy s javascriptem v některých stránkách
• vývoj ukončen, přechod na Wayback

2.4.6 Wayback

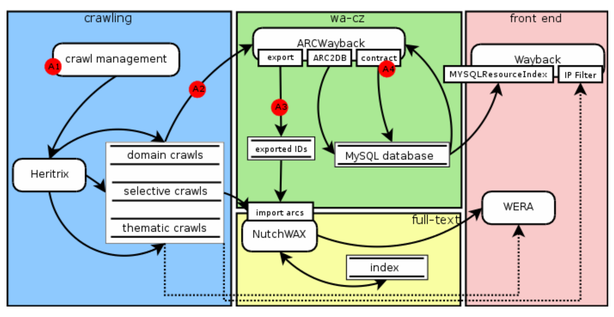
• aplikace pro zpřístupnění archivu, která v budoucnu nahradí stávající Wayback Machine Internet Archivu
• dokumenty jsou indexovány a zpřístupňovány pomocí URL a času
• režimy zpřístupnění:
  • archival URL = úprava odkazů na stránce (link zpět do archivu)
  • proxy = chová se jako proxy server, ale je pak složité měnit časové verze (WAX Toolbar – plugin pro Firefox)
  • timeline = časová osa, zatím experimentální
• připravuje se podpora fulltextového vyhledávání
• dokončena lokalizace verze 0.9 do češtiny

3 Implementace v rámci projektu WebArchiv

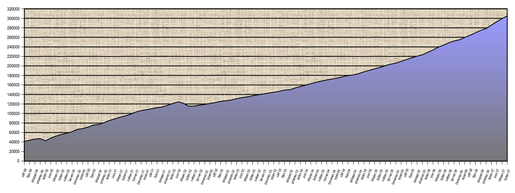
- stahování Webu pomocí Heritrixu
- neomezené online zpřístupnění části archivu pomocí WERA (včetně fulltextu)
- zpřístupnění clého archivu pomocí WayBack (jen URL, přístup k dokumentům jen z budovy knihovny)
- vlastní vývoj – databáze obsahující informace o všech souborech v archivu a její integrace s nástroji IIPC
- příprava implementace OAI-PMH

4. Statistika

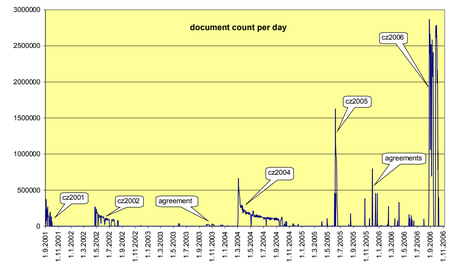
- počet sklizených dokumentů ke dni 8.5.2007 je 134,5 miliónů
- objem sklizených dat je 5 465 GB
- první dokument byl archivován 3.9.2001
- proběhla analýza zaměřená na složení archivu podle typů souborů a jejich velikostí, zaměřená na možnosti úspory úložného prostoru



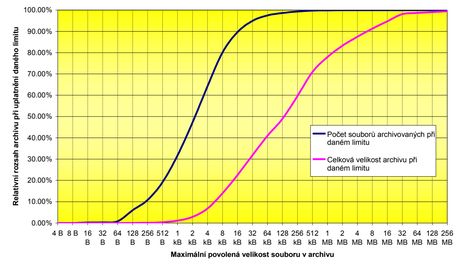
A1 *nová sklizeň*
A2 *konec sklízení -> indexovat*
A3 *aktualizovat fulltext*
A4 *aktualizovat seznam souborů*



Počet registrovaných domén v zóně .cz vzrostl od roku 1999 z 40.000 na více než 300.000



Přírůstky souborů WebArchivu v čase



Srovnání počtu souborů a velikosti archivu při zavedení limitu na velikost souborů

http://www.webarchiv.cz/