Vyhledávání v archivu českých webových zdrojů

Mgr. Jan HUTAŘ

Bc. Lukáš MATĚJKA

Mgr. Ludmila CELBOVÁ

Proč vznikl WebArchiv?

- archivace elektronických online zdrojů je celosvětovým trendem
- Potřeba zachránit netištěné informace kulturní a historické hodnoty pro další generace
- až 90% webových dokumentů existuje pouze v elektronické podobě (periodika, monografie, konferenční příspěvky, výzkumné a jiné zprávy, akademické práce, WWW stránky, dokumenty státní správy atd.)
- NK ČR je *depozitní knihovnou*, odpovídá za trvalé uchovávání fondu bohemikálních dokumentů jako součásti národního historického a kulturního dědictví

Jak vznikl WebArchiv?

- WA vznikl v rámci programového projektu MK ČR *VaV - "Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet*"
- řešen od roku **2000** v NK ČR ve spolupráci s MZK Brno a ÚVT Masarykovy univerzity v Brně
- pouze díky menším grantovým podporám se daří projekt udržet i nadále rozvíjet

Cíle WebArchivu

- zajistit pokud možno *trvalý přístup* k „domácím" elektronickým zdrojům publikovaným v síti Internet ✓
- připravit podmínky pro získávání, zpracování, archivaci a ochranu online přístupných elektronických zdrojů ✓
- zajistit zpřístupnění zdrojů z digitálního archivu za podmínek respektujících autorské právo ✓
- stanovit kritéria výběru zdrojů pro národní bibliografii ≈
- zajistit technické a programové řešení indexace a archivace elektronických online zdrojů ≈
- zajistit, implementovat a udržovat standardy pro budoucí čitelnost zdrojů a pro vyhledávání v síti ×
- vytvořit podmínky pro kooperaci knihoven (inf. pracovišť) a propojení s vydavateli elektronických zdrojů ×

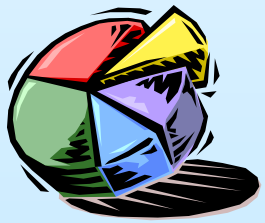
Kritéria výběru webových zdrojů

- množství online dokumentů je obrovské, jejich kvalita různá → snaha uchovat dokumenty, které mají dokumentární hodnotu pro současné i budoucí generace → *tj. aplikovat kritéria výběru*

Pro akvizici zdrojů se aplikují dva přístupy:

1. ***výběrová archivace -*** *sklízejí* a archivují se pouze dokumenty vybrané podle určitých kritérií
2. ***plošná archivace (harvesting)* –** např. celé národní domény. Nutná pouze kritéria technické povahy a nastavení harvesteru.

- přístupy v jednotlivých zemích různé
- trend – oba přístupy najednou (např. Austrálie, Dánsko)



Kritéria výběrové archivace

- stanovení kritérií bylo velmi komplikované a stále pokračuje, koordinujeme projekt "Web Cultural Heritage" v rámci EU Culture 2000

- **Obsah** - odborné, umělecké či zpravodajsko -publicistické webové zdroje – *nejpodstatnější kritérium*
- **Národní aspekt** - „národnost autora", „národní jazyk", „země nebo národ jako téma"
- **Doména** - .cz (com, net apod.)
- **Přístup** - zdroj volně přístupný nebo pod heslem
- **Formát** - zdroje ve formátech interpretovatelných běžným internetovým prohlížečem
- **Původní forma zdroje** – zdroje mající pouze! elektronickou verzi
- **Typ zdroje** - např. konferenční materiály, monografie, online seriály, akademické práce apod.

Co máme za sebou

- průběžné testování:
  - SW nástrojů s využitím HW pořízeného v rámci finančních možností
  - tj. aplikací pro *stahování, archivaci, indexaci a zpřístupnění* webových stránek
- SW výhradně open source

- snaha o změnu zákonů
- mezinárodní spolupráce (aktivní účast na výzkumu a vývoji v rámci IIPC – ač nejsme zatím členy)
- zpřístupňování veřejné části archivu od podzimu 2005
- zpřístupnění celého archivu během 1 měsíce (lokálně)

Provedené sklizně domény .cz

- **2001** první pokus o testovací celoplošnou sklizeň domény .cz, 1 stroj + páskový robot, cz2001 obsahuje přes 3 mil. jedinečných URL (107 GB) – nedokončena z tech. důvodů
- **2002** harvester sklízel několik měsíců → přerušeno pro omezený výkon serveru a záplavy. cz2002 obsahuje 315,5 GB, z 10 263 855 URL staženo přes 10 mil. dokumentů (→ tematická sklizeň Povodně)
- **2004** 03/10 (v roce **2003** neproběhla), z 32 149 396 URL staženo 32,5 milionu souborů = 1,2 TB. Nový server a nové diskové pole
- všechny sklizně prováděny s **NEDLIB** harvesterem, hloubka zanoření 25-50 odkazů
- od roku 2004 nový harvester **HERITRIX** - několik sklizní hlavních stránek většiny českých domén

Současný stav projektu

- **6x ročně** sklízen soubor zdrojů (asi 110 serverů), na které má NK smlouvu o zpřístupnění. Přírůstek okolo 10GB komprimovaných dat /rok → zvyšuje se
- sklízení omezených množin webových zdrojů je úspěšné
- **ALE** … od roku 2004 se nedaří dlouhodobě udržet v provozu souvislou sklizeň domény .cz - problém Heritrixu s využitím paměti
- nová verze Heritrixu má tento problém odstranit → **letos** počítáme se spuštěním celoplošné sklizně domény .cz
- při přípravě tzv. semínek (startovací URL Heritrixu) provedena analýza domény .cz → vyřazeny servery podezřelé z nerelevantního obsahu (mail, mysql apod.) a duplicitní (http://www.centrum.cz a http://centrum.cz)

Současný stav projektu pokračování

- v současné době je ve **WebArchivu** uloženo asi 1,7 TB dat ≈ 50 milionů archivovaných unikátních dokumentů
- snaha o sklizení celé domény .cz 1x ročně, spíše 2x
- zdroje na které máme uzavřeny s vydavateli smlouvy pro zpřístupňování jsou sklízeny 4x do roka.
- konec 2006 → vše uloženo na nové digitální úložiště dat NK ČR
- 2007 archiv projektu by se měl stát součástí připravované „Digitální knihovny" NK ČR

Změny softwarového vybavení

- 2004 vývoj a podpora NEDLIB harvesteru definitivně zastavena
- **2004-2005** postupný přechod na SW vyvíjený konsorciem IIPC (International Internet Preservation Consortium)
- Webarchiv Indexer (MFF UK) nahrazen NWA toolsetem (využívá index vytvořený Apache Lucene)
- archivní souborový formát **nedlib** nahrazen **ARC** formátem (používá jej Heritrix) → nutno převést již uložená data na nový ARC formát. Vytvořen nástroj NedlibToArc v rámci NWA, obsahuje chyby (zvláště při převodu velkých souborů)

Heritrix – výhody

modulárnost, rozšiřitelnost, **neustálý vývoj** (v.1.6), kvalitní a rychlá **podpora** vývojářů z IA

open source kódy a modularita umožňují spolupráci třetích stran na jeho vývoji

- 2 základní části - **framework** a **přípojné moduly**
  - *Framework* - základní kontrola nad sklizněmi, uživatelské rozhraní, správu běžících procesů a nastavení sklizní
  - *přípojné moduly* - použity pro konkrétní implementaci sklizní, určují každý krok sklizně

Heritrix - problémy

- nelze dlouhodobě sklízet web bez odborných zásahů
- získávání odkazů z webových stránek
  - extrahování linků z entit href, src, img je snadné, ale ne všechny linky se vyskytují v HTML značkách (JavaScript)
- tzv. detekce pastí
  - vkládání sessionID při generování stránek → nutno individuálně ošetřit takové servery + filtry Heritrixu
- využití paměti při celoplošných sklizních
  - prozatímní řešení: rozdělit počáteční semínka na skupiny
- inkrementální sklízení a detekce změn
  - nelze individuálně díky obrovskému počtu zkoumaných stránek.
- hloubka sklizně, počet sklízených dokumentů z 1 serveru, délka sklizně… → nutné hledat kompromisy

SW pro zpřístupnění

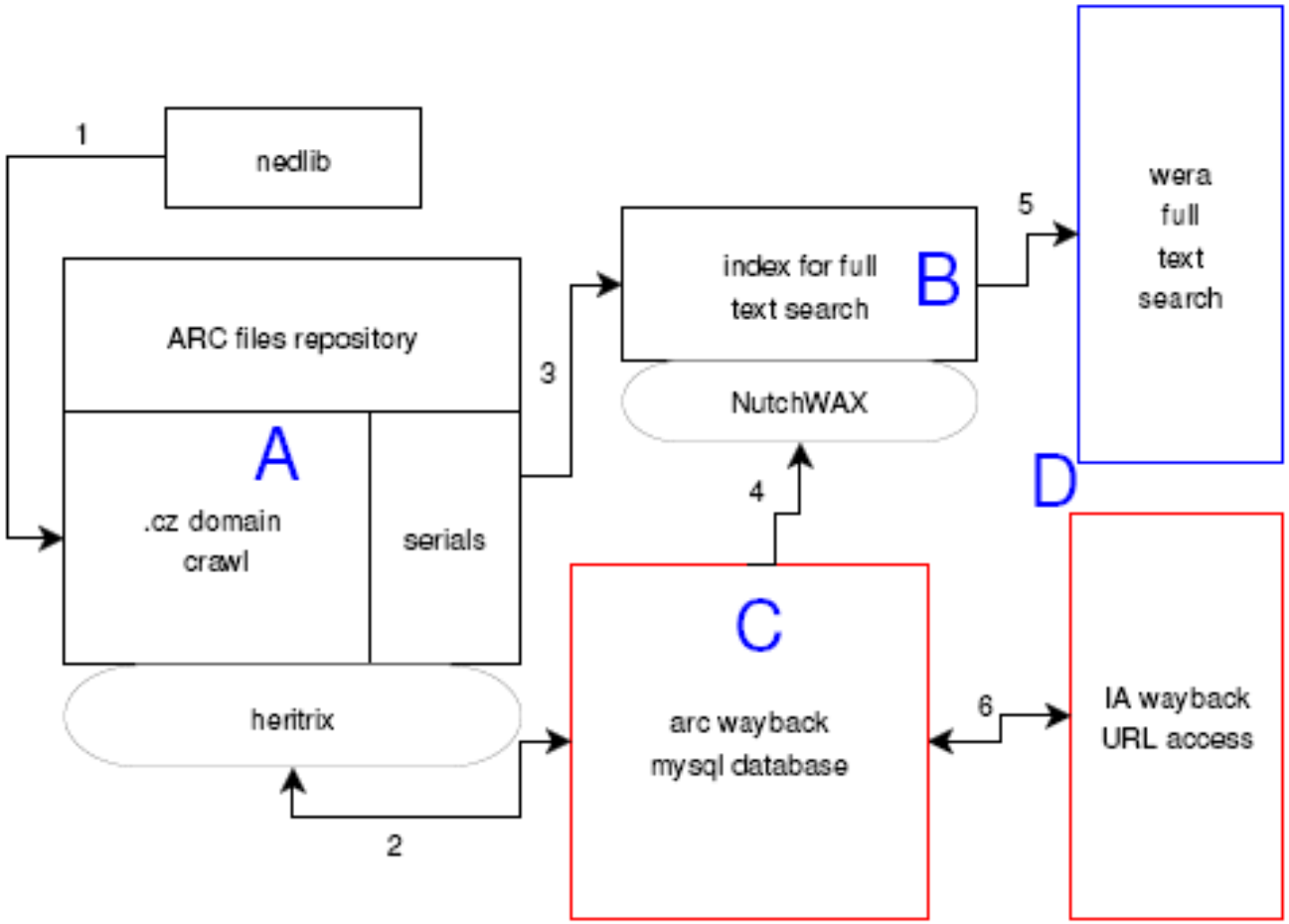
- výrazný kvalitativní posun
- fulltextové indexování dokumentů - systém **NutchWAX**, nástavba nad vyhledávacím rozhraním **Nutch**
- **WERA** (následník NWA tools) - uživatelské rozhraní pro zobrazování dokumentů (vyhledávání v archivu přes www rozhraní), využívá index NutchWAXe. Zvládá českou diakritiku – vyhledávání v indexu a zobrazování českých dokumentů
- vytvořit fulltextový index nad celým archivem je velmi technicky a výpočetně náročné → bylo nutné sestrojit jiný index umožňující *přístup do archivu podle URL a času*

Návrh a realizace systému

- **archivace dokumentu** – pomocí crawleru Heritrix jsou dokumenty ukládány na datové úložiště ve formátu ARC
- užitečný archiv = přístup k archivovaným dokumentům → pro tyto účely nutné zkonstruovat index pro zpřístupnění a k němu rozhraní, které zobrazí výsledky
- vytvoření fulltextového **indexu nad kolekcí** výběrových zdrojů pro vyhledávání podle slov - nástroj NutchWAX
- sestrojení **globálního indexu** pro zpřístupnění celého archivu
- **zpětné zobrazení dokumentů** z archivu – nástroj WERA a Wayback

Návrh a realizace systému pokračování

- vlastností definice archivního formátu ARC je, že soubor lze identifikovat a extrahovat bez dalších souvisejících indexů, které by tyto soubory popisovaly. Tj. pokud není vytvořen příslušný index → v souborech lze vyhledávat jen sekvenčním přístupem = dlouhý přístup k datům
- NutchWAX přístup urychlí - vytvoří fulltextový index
- nemůžeme vytvořit tak rozsáhlý index pro celý archiv (výpočetně složité a náročné datové úložiště) → NutchWAXem indexujeme jen omezené množiny dokumentů (výběrové zdroje se smlouvou)
- pro zbytek archivu potřeba nástroj, který jej zpřístupní. Součástí aplikace by měl být méně rozsáhlý index, který efektivně vyhledává dokumenty na základě URL a času


1
nedlib
ARC files repository
A
.cz domain
crawl
serials
heritrix
2
3
index for full
text search
B
NutchWAX
4
5
wera
full
text
search
D
C
arc wayback
mysql database
6
IA wayback
URL access

Nástroje využívané pro zpřístupnění

- Nutch
- NWA
- WERA
- ArcRetriever
- NutchWAX
- Wayback

Nutch

- volně šířitelné transparentní **vyhledávací rozhraní (engine)** implementované plně v Javě
- má plnou škálu nástrojů, které jsou třeba k provozování vlastního vyhledávacího enginu (např. vlastní crawler)

Nutch dokáže:

1. stáhnout a zpracovat miliony stránek měsíčně
2. spravovat index těchto stránek
3. vyhledávat v takovém indexu 1000x za vteřinu a poskytovat velmi kvalitní výsledky vyhledávání

- vhodný pro různorodé prostředí (intranet i Internet)
- vychází z architektury Apache Lucene

WERA - WEb aRchive Access

- vznikla ze spolupráce konsorcia IIPC, Internet Archive a severských zemí
- využívá hlavní části NWA (prohlížeč archivních dokumentů)
- předností je velmi snadná navigace a propracované uživatelské rozhraní (časová osa zobrazuje časové verze dokumentu)
- výsledky vyhledávání v podobě URL zobrazeny velmi přehledně a u každého odkazu jsou linky na získání dalších časových verzí téhož URL
- zobrazovat archivované stránky lze i pomocí zadání přesné URL adresy
- archivované dokumenty a WERA propojeny skrz index Nutche

NutchWAX - NutchWeb Archive Extensions

- **nástavba vyhledávacího rozhraní Nutch** vytvořená pro potřeby indexování dokumentů archivovaných Heritrixem
- je to soubor indexovacích a dotazovacích pluginů, které přidávají do indexu potřebná metadata
  - např. jméno archivního souboru, ve kterém je dané URL uloženo, jméno kolekce apod.
- vlastní zpracování ARC souborů – pomocí modulu arc2seg konvertovány do nutch-segmentů a ty jsou dále indexovány

Wayback

- prototypní SW – java aplikace
- měl by časem nahradit stávající Wayback Machine
- indexační modul - průběžně spravuje index s URL záznamy nad archivními soubory ARC
- uživatelské rozhraní umožňuje zobrazovat dokumenty na základě URL a času a také pomocí tzv. **hvězdičkové konvence** (vyhledávání pomocí levých prefixů URL)

- aplikace je v neustálém vývoji
- počítá se i s fulltextovým prohledáváním

Budoucnost projektu

- **hlavní cíl** – 2006 spustit a dlouhodobě udržet celoplošnou sklizeň domény .cz
- nadále sbírat smluvní výběrové zdroje, jejichž význam se stále zvětšuje
- zautomatizovat proces od výběru zdroje, oslovení vydavatele k podpisu smlouvy s vydavatelem → fulltextové vyhledávání v archivu
- legální lokální zpřístupnění celého archivu (vyhledávání podle URL a času sklizně dokumentu) – během 1 měsíce
- zavedení tzv. inkrementální sklízení (identifikace změn v opakovaně sklízených dokumentech)

Budoucnost projektu pokr.

- rozeznání duplicitních dokumentů (budoucí verze crawleru)
- inkrementální indexování - přidávání nových dokumentů do již existujícího indexu, tj. nevytváření nového
- identifikace domácích dokumentů mimo doménu .cz – procházení stránek nástrojem pro identifikaci jazyka
- fulltextové indexování celého archivu – proces nutno distribuovat na několik strojů (další release NutchWAXu)
- jednoznačný identifikátor dokumentů v archivu
- podpora OAI protokolu
- 2007 integrace do připravované „Digitální knihovny ČR”