Vyhledávání v archivu českých webových zdrojů

Mgr. Jan HUTAŘ
Národní knihovna ČR
jan.hutar@nkp.cz

Bc. Lukáš MATĚJKA
Fakulta informatiky MU Brno
lukas.matejka@centrum.cz

Mgr. Ludmila CELBOVÁ
Národní knihovna ČR
ludmila.celbova@nkp.cz

INFORUM 2006: 12. konference o profesionálních informačních zdrojích

Praha, 23. - 25.5. 2006

Abtrakt. *První informace o projektu WebArchiv zaměřeném na archivaci českých elektronických zdrojů publikovaných v síti Internet byla prezentována na konferenci INFORUM v roce 2000. V roce 2004 byl problematice získávání a archivace elektronických zdrojů, zejména zdrojů internetových věnován workshop. Po pěti letech řešení projektu nazvaného WebArchiv umožnili řešitelé přístup do části webového archivu.*
Příspěvek seznámí s technickým zázemím, objasní jak byl archiv zpřístupněn, co lze v archivu nalézt a co můžeme v blízké budoucnosti očekávat ve zlepšení možností zpřístupňování archivovaných webových zdrojů.
Druhá část příspěvku se zabývá vývojem a aplikací technického řešení procesu zpřístupnění (implementace SW, HW).

PROČ VZNIKL WEBARCHIV?

Archivace elektronických online zdrojů je celosvětovým trendem. Neexistuje jiná cesta, jak zachránit tyto netištěné informace kulturní a historické hodnoty pro další generace, než cesta archivace. Proto se o to snaží i Národní knihovna ČR, která je *depozitní knihovnou*, odpovědnou za trvalé uchovávání fondu bohemikálních dokumentů jako součásti národního historického a kulturního dědictví. Tyto dokumenty jsou uchovávány v národním konzervačním fondu. Dosud v něm byly uchovávány a v České národní bibliografii registrovány *pouze* klasické dokumenty (tištěné, zvukové atd.).
Cílem projektu **WebArchiv** je zajistit trvalý (dlouhodobý) přístup také k "domácím" *elektronickým* zdrojům publikovaným v síti Internet. Až 90 % z těchto dokumentů existuje pouze v elektronické podobě (periodika, monografie, konferenční příspěvky, výzkumné a jiné zprávy, akademické práce, WWW stránky, dokumenty státní správy atd.).

JAK VZNIKL WEBARCHIV?

WebArchiv je projekt, který zastřešuje snahu o *dlouhodobé uchování a zpřístupnění online dostupných elektronických informačních zdrojů*. Projekt vznikl v rámci programového projektu výzkumu a vývoje "*Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet*". Byl řešen od roku 2000 v Národní knihovně ČR za částečné grantové podpory Ministerstva kultury ČR, ve spolupráci s Moravskou zemskou knihovnou v Brně a Ústavem výpočetní techniky Masarykovy univerzity v Brně. Díky menším grantovým podporám Ministerstva kultury ČR se daří projekt i nadále rozvíjet a postupně připravovat podmínky pro uvádění výsledků výzkumu do každodenní praxe.

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 centrálních, regionálních a specializovaných knihoven, resp. informačních pracovišť a propojení s vydavateli elektronických zdrojů

Archivace webu je velmi komplexní činnost náročná v každém ze svých aspektů. Zmíníme se proto jen o některých.

KRITÉRIA VÝBĚRU WEBOVÝCH ZDROJŮ

Vzhledem k tomu, že množství dokumentů přístupných online je obrovské a publikace zveřejňované na Internetu jsou velmi rozdílné kvality, je třeba pro účely tvorby archivu webových zdrojů aplikovat určitá kritéria výběru tak, aby byly uchovávány dokumenty, jež mají dokumentární hodnotu pro současné i budoucí generace.
Technicko-knihovnická kritéria, podle nichž jsou vybírány webové zdroje určené pro uložení v archivu a pro zpracování do České národní bibliografie (ČNB), byla stanovena na základě zkušeností s dosavadním řešením projektu **WebArchiv**. Pro akvizici zdrojů se aplikují dva přístupy:

- ***výběrová archivace (s vyšším podílem intelektuální práce)***, kdy se „*sklízejí*“[1] a archivují pouze dokumenty vybrané podle určitých kritérií (viz dále)
- ***plošná archivace (převážně automatický proces - harvesting)***, kdy se sklízí např. celá národní doména, (u nás tedy doména .cz). Při plošné sklizni je nutné stanovit pouze kritéria technické povahy a na základě zkušeností nastavit harvester (open source SW Heritrix). Může se jednat o omezení maximálního počtu dokumentů sklizených z jedné webové stránky, omezení maximální velikosti sklízeného dokumentu (např. 100 MB), podporované protokoly apod.

V ČR (a projektu WebArchiv) se využívají oba přístupy, tedy plošná sklizeň i výběrová archivace. V jednotlivých zemích se pohledy na to, zda využívat výběr nebo celoplošnou sklizeň příslušného národního webu, liší. Můžeme ovšem říci, že tam kde se využíval pouze výběrový přístup, později také sáhli k celoplošné sklizni (např. Austrálie[2], Dánsko[3]).

Kritéria výběrové archivace

Pro účely registrace v ČNB a souběžné uchování v digitálním archivu je důležité vybírat zdroje významné z hlediska kulturně historického. K tomu slouží kritéria výběru. Proces ustanovení těchto kritérií byl velmi komplikovaný a na jejich upřesnění se stále pracuje. Nejnověji pak v rámci projektu Web Cultural Heritage[4] (program EU Culture2000), jehož je Národní knihovna ČR koordinátorem. O kritériích by se dalo psát velmi obšírně, ovšem bylo by to mimo téma tohoto článku.
Jedním z nejdůležitějších je kritérium **„Obsah“**. Mělo by jít o webové zdroje odborného, uměleckého či zpravodajsko-publicistického zaměření. Dále se jedná o **„Národní aspekt“** dané země: „národnost autora“, „národní jazyk“, „země nebo národ jako téma“, což odpovídá kriteriím pro klasickou národní bibliografii. Dále jsou kriterii **„Doména“** (cz, možno uvažovat i com, net apod.), **„Přístup“** (zdroj volně přístupný nebo pod heslem), **„Formát“** (zdroje ve formátech interpretovatelných běžným internetovým

[1] takto se označuje proces stahování dat z prostředí Internetu (z angl. „harvesting“)
[2] KOERBIN, Paul. Report on the Crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005 [online]. [cit. 18.4.2006]. Přístup z WWW : <http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf>.
[3] ANDERSEN, Bjarne. The DKdomain: in words and figures [online]. [cit. 18.4.2006]. Přístup z WWW : <http://netarkivet.dk/publikationer/DFrevy_english.pdf>
[4] Web Cultural Heritage [online]. [cit. 18.4.2006]. Přístup z WWW : <http://www.webarchiv.cz/culture-2000/>.

prohlížečem), **„Původní forma zdroje“** (zdroje, které nemají tištěnou verzi, jsou tzv. digital born) a konečně **„Typ zdroje“** (např. konferenční materiály, monografie, online seriály, akademické práce apod.). Nutno říci, že vždy záleží na výběru té které země a na rozhodnutí, jak bude „své“ zdroje do archivu vybírat. Podrobnosti viz [5].

CO MÁME ZA SEBOU

Oblast IT k projektu **WebArchiv** zajišťuje externě spolupracující Ústav výpočetní techniky Masarykovy univerzity v Brně. Průběžně provádí testování SW nástrojů s využitím HW pořízeného v rámci finančních možností. Jedná se zejména o aplikace pro stahování, archivaci a indexaci/zpřístupnění webových stránek (viz dále).
V rámci pilotního projektu proběhl v roce 2001 první pokus o testovací celoplošnou sklizeň (harvest) domény .cz. V tehdejších podmínkách s jedním strojem a jedním úložištěm v podobě páskového robota. Sklizeň nazvaná cz2001 obsahuje přes 3 miliony jedinečných URL a zabírá téměř 107 GB nekomprimovaných dat. Přestože sklízení domény nemohlo být z technických příčin dokončeno, získané zkušenosti umožnily připravit se lépe na další sklizeň, která následovala v roce 2002. Nedlib harvester po dobu několika měsíců sbíral data, až do té doby kdy byla sklizeň nuceně přerušena. Sklizeň nazvaná cz2002 obsahuje 315,5 GB nekomprimovaných dat. V jejím rámci bylo alespoň jednou navštíveno přes 33.000 domén druhé úrovně (zhruba jedna čtvrtina tehdejšího počtu v doméně .cz), z 10 263 855 URL bylo staženo přes 10 milionů dokumentů. Pro omezený výkon sklízecího serveru a také kvůli srpnovým záplavám nemohla být tato sklizeň dokončena. Na náhradním HW byla provedena alespoň malá tematická sklizeň zaměřená na povodňové zpravodajství.
V rámci sklizně březen - říjen 2004 (v roce 2003 sklizeň neproběhla) bylo z 32 149 396 URL staženo 32,5 milionu souborů a byl tak vytvořen archiv o celkové velikosti 1,2 TB (po kompresi 611 GB). V témže roce byl zakoupen nový server a nové diskové pole pro uložení dat.
Všechny tyto sklizně byly prováděny pomocí programu NEDLIB Harvester při hloubce zanoření až 25-50 odkazů.
Od poloviny roku 2004 pak bylo provedeno několik sklizní hlavních stránek většiny českých domén pomocí nového harvesteru Heritrix (viz dále).

SOUČASNÝ STAV PROJEKTU

V současnosti se stále využívá crawler Heritrix. V průběhu roku se zhruba šestkrát sklízí soubor zdrojů čítající desítky serverů[6], na které má NK smlouvu o zpřístupnění. Přírůstek objemu dat se pohybuje průměrně kolem 10GB komprimovaných dat a stále se zvyšuje s přibývajícími vydavateli.
Na rozdíl od poměrně úspěšného sklízení omezených množin webových zdrojů se od roku 2004 nepodařilo dlouhodobě udržet v provozu souvislou sklizeň domény .cz, a to díky problémům Heritrixu s využitím paměti. Heritrix totiž obvykle již po několika dnech provozu spotřeboval všechnu dostupnou paměť díky velkému množství odkazů, které se chystal „navštívit“. Současná verze Heritrixu slibuje tento problém odstranit, takže se počítá se spuštěním celoplošné sklizně domény .cz již v tomto roce.
K přípravě tzv. semínek (startovací URL pro Heritrix) pro celoplošnou sklizeň byla provedena analýza na doméně .cz, jejíž seznam byl zakoupen ze serveru nic.cz. Byly tak vyřazeny servery, které se zdály „podezřelé“ z nerelevantního obsahu (např. názvy začínající mail, mysql, user apod.) nebo byly duplicitní (např. http://www.centrum.cz a http://centrum.cz). Celkový počet URL tedy klesl na asi 378 tisíc z původních 540 tisíc.

V současné době je ve **WebArchivu** uloženo asi 1,7 TB dat, což představuje zhruba 50 milionů archivovaných unikátních dokumentů. Snahou je, aby sklizně celé domény .cz probíhaly pokud možno jednou ročně, zdroje na které máme uzavřeny s vydavateli smlouvy pro zpřístupňování jsou sklízeny přibližně čtyřikrát do roka.
Počet sklizní je do jisté míry limitován výkonem serverů, kapacitou úložného prostoru a funkčností používaného softwaru, který se průběžně vyvíjí. Koncem letošního roku by všechna data měla být uložena na novém digitálním úložišti dat Národní knihovny ČR, které by mělo i do budoucna zaručovat dostatek úložného prostoru pro další rozšiřování archivu. Zároveň by se archiv projektu WebArchiv měl v roce 2007 stát součástí připravované „Digitální knihovny“ v Národní knihovně ČR.

[5] Podrobnosti o posledním vývoji selekčních kritérií viz http://www.webarchiv.cz/culture-2000-documents/
[6] aktuální seznam spolupracujících vydavatelů viz http://www.webarchiv.cz

LEGISLATIVA

Současná legislativa neumožňuje či zpochybňuje oprávnění depozitních knihoven vytvářet „konzervační“ sbírku v digitálním archivu a zcela vylučuje možnost tyto dokumenty dále veřejně zpřístupňovat! Pro potřeby tohoto článku postačí uvést pouze základní skutečnosti.
V České republice neexistuje povinný výtisk online publikací, tj. právo knihoven na povinný výtisk elektronických online dokumentů není ve stávajících zákonech o povinném výtisku zaneseno, což práci na projektu velmi ztěžuje. Je nutno jednat jednotlivě s každým vydavatelem a vyžadovat souhlas ke zpřístupnění jeho zdroje skrz rozhraní WebArchivu, i když zdroj je sklizen a uložen v rámci celoplošné sklizně.
Národní knihovna ČR, která má za úkol uchovávat národní kulturní dědictví, usiluje proto o novelizaci potřebné legislativy. Příklady podobných opatření nalezneme v ostatních zemích (např. Velká Británie, Německo, Rakousko, Francie, Švédsko či Finsko).

Podobná situace je i okolo autorského zákona, kde stojí proti sobě požadavky knihoven (umožnit uživatelům přístup k online zdrojům a pořizování kopií) a požadavek vydavatelů/autorů (předejít zneužití autorských děl jejich uživateli, tj. neoprávněné využívání, kopírování). Je potřeba nalézt rovnováhu mezi oběma postoji. V současné době se čeká na novelu autorského zákona[7], která má umožnit zpřístupňovat webový archiv (nebo jinou elektronickou sbírku) alespoň lokálně, tzn. v budově knihovny, přesněji instituce, kde je archiv/sbírka uložen, a to na speciálních terminálech bez možnosti kopírování prohlíženého obsahu a výhradně ke studijním účelům.
Bez možnosti zpřístupnění celého archivu, v podstatě účel WebArchivu pozbývá smyslu.

NÁHRADNÍ ŘEŠENÍ

Z výše uvedených legislativních důvodů jsme již na začátku projektu přistoupili k **oslovování jednotlivých vydavatelů a uzavírání smluv**[8] o poskytování elektronických online zdrojů. **Není to však řešení ideální**, poněvadž je časově velmi náročné. V rámci našich personálních kapacit je možné tímto způsobem oslovit několik stovek, možná tisíc vydavatelů.

ZMĚNY SOFTWAROVÉHO VYBAVENÍ

Mezi lety 2004-2005 se postupně přecházelo na SW vyvíjený konsorciem IIPC (International Internet Preservation Consortium). Nedlib harvester, jehož vývoj byl definitivně zastaven, byl nahrazen crawlerem Heritrix. Pro zpřístupnění se místo Webarchiv Indexeru[9] začal používat NWA toolset, který používal index vytvořený Apache Lucene. Změna nástrojů vedla i ke změně archivního souborového formátu z nedlib na ARC formát používaný Heritrixem. Z toho vyplynula potřeba převést data ve starém formátu nedlib na nový formát. K tomu byl vytvořen nástroj NedlibToArc v rámci NWA, který taktéž obsahuje několik chyb (zvláště při převodu velkých souborů).

NÁSTROJE PRO ARCHIVACI – HERITRIX

Heritrix je open-source software[10] vyvíjený společností Internet Archive. Mezi hlavní výhody patří modulárnost a rozšiřitelnost. Volně dostupné kódy - Open-source a modularita programu umožňuje, aby se na vývoji a přizpůsobení Heritrixu na speciální podmínky podílely i další strany. Heritrix je rozdělen do dvou základních částí - framework a přípojné moduly. Framework zajišťuje základní kontrolu nad sklizněmi, uživatelské rozhraní, správu běžících procesů a nastavení sklizní. Pro konkrétní implementaci sklizní jsou použity přípojné moduly, které určují každý krok sklizně. Heritrix obsahuje mnoho implementací těchto modulů, které umožňují spouštět velmi rozsáhlé sklizně. Další výhodou je, že Heritrix je v neustálém vývoji (nyní v. 1.6). Za zmínku stojí velmi kvalitní podpora vývojářů Internet Archive. Slabinou systému je nemožnost dlouhodobě sklízet web bez odborných zásahů. S rostoucím počtem zachycených URI Heritrix „padá“ na problémech s pamětí. Dalším

[7] z. 21/2000 Sb.

[8] podrobnosti v prezentaci viz http://www.webarchiv.cz/files/dokumenty/seminar/Infoden_061205.ppt

[9] nástroj Webarchiv Indexer pro zpřístupnění dat v archivu, vznikl jako roční projekt studentů MFF UK a uveden do provozu byl v roce 2004, kdy zpřístupňoval část archivu (kolekci zvanou serials, tj. ty zdroje, na které měla NK s vydavateli uzavřeny smlouvy). Indexer obsahoval závažné chyby a jeho vývoj dále nepokračoval.

[10] oficiální web Heritrixu a možnost stažení na http://crawler.archive.org

problémem, který se momentálně usilovně řeší, je nemožnost inkrementálního indexování na celoplošných sklizních.

Heritrix a proces sklízení

Pro akvizici (získání) obsahu Internetu se v současnosti v ČR používá systém Heritrix. Při používání tohoto nástroje se vyskytují určité nedokonalosti, které jsou ovšem řešeny v následujících verzích softwaru.
Jedním ze zásadních problémů je získávání odkazů z webových stránek. Extrahování linků z entit href, src, img je poměrně snadné, avšak ne všechny linky se vyskytují v těchto HTML značkách[11]. Automatické sklizně se potýkají s problematikou tzv. detekce pastí. Jde o dynamicky generované dokumenty, které se navenek tváří jako zdroj velkého množství unikátních dokumentů, ačkoliv nejde o změnu informačního obsahu. Příkladem takového serveru, který je sklízen v rámci nasmlouvaných zdrojů, je militaria.cz. Zde dochází při generování stránek ke vkládání tzv. sessionID, což je speciální parametr, který označuje „sezení" pro uživatele. Pro crawler je to problém, protože identifikuje URL s odlišným sessionID jako nové unikátní URL. Bohužel rozpoznání takových pastí je velmi obtížné, zvláště při celoplošných sklizních, kde se stahuje obrovské množství dat. Je tedy nutné individuálně ošetřit „problémové" servery a nastavit filtry v Heritrixu, které eliminují toto nežádoucí chování.
Heritrix se v současnosti při celoplošných sklizních potýká s problémy využití paměti. Vzhledem k tomu, že crawler musí projít velmi rozsáhlou frontu linků, které jsou odkazovány z tzv. „semínek", dochází k přeplnění paměti. Prozatímní řešení je rozdělit počáteční semínka na skupiny (např. podle písmen abecedy) a sklízet tuto omezenou množinu jednou instancí Heritrixu. Toto má ovšem zásadní nevýhodu - není možné plně využít vzájemných odkazů mezi takto vzniklými částmi archivu.
Uchování digitálního obsahu Internetu si vyžaduje poměrně často postřehnout změny na webových stránkách. Tuto otázku lze řešit inkrementálním sklízením. Díky obrovskému počtu zkoumaných stránek je velmi složité nějakým způsobem zajistit časté sledování změn v dokumentech a dle toho rozhodovat o jejich opětovném sklízení. Některé stránky se mění dynamicky velmi rychle, aniž by se podstatně změnil jejich obsah.
Mezi další otázky patří zvolení optimální hloubky sklizně, počet sklízených dokumentů v rámci serveru, délka sklizně[12] apod. Je nutné hledat kompromisy, které napomohou k co možná nejvíce informačně hodnotné sklizni dokumentů z webu.

ZPŘÍSTUPNĚNÍ WEBARCHIVU

WebArchiv jako takový může v současné situaci zpřístupňovat veřejně pouze tu část, lépe řečeno ty zdroje, na které má uzavřenu smlouvu s jednotlivými vydavateli. Pokud bude schválena novela AZ, zejména §37 týkající se zpřístupňování rozmnoženin děl, budou všechny archivované dokumenty uložené v našem digitálním archivu zpřístupněny na určených terminálech lokálně a vyhledávání bude možné na základě URL a času (viz dále).

Změna softwaru pro zpřístupnění

V oblasti použitých nástrojů pro zpřístupnění došlo k výraznému kvalitativnímu posunu. K fulltextovému indexování dokumentů se začal používat systém NutchWAX, což je nástavba nad vyhledávacím rozhraním Nutch. WERA, následník NWA tools, slouží jako uživatelské rozhraní pro zobrazování dokumentů (vyhledávání v archivu přes www rozhraní) a využívá index NutchWAXe. Po společném úsilí spolu s vývojáři Internet Archive se podařilo odstranit chyby a v současnosti tyto nástroje dokáží dobře zacházet s českou diakritikou, vyhledávat v indexu a zobrazovat správně kódované dokumenty.
Vytvořit fulltextový index nad celým archivem je velmi technicky a výpočetně náročné, bylo proto nutné sestrojit jiný index, který by umožňoval přístup do archivu podle URL a času.

Návrh a realizace systému

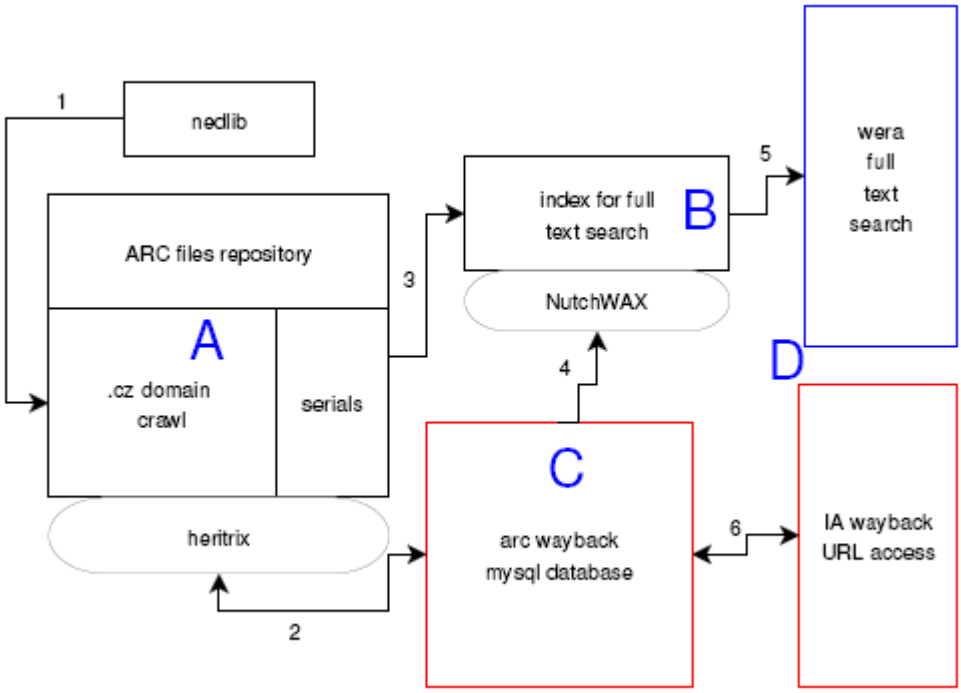
Proces archivace a zpřístupnění elektronických dokumentů se skládá z mnoha nesnadných kroků.
Celý systém vychází z několikaletých zkušeností projektu WebArchiv. Postup zpracování dokumentu můžeme rozdělit na následující procesy:

[11] např. konstrukce linků pomocí JavaScriptu, nebo extrahovaní linků z jiných formátů jako PDF, MSWord, Flash
[12] sklizeň může běžet až do vyčerpání HW prostředků, je tedy nutné ji nějak časově omezit

- archivace dokumentu – pomocí dostupného nástroje (crawleru Heritrix) jsou dokumenty ukládány na datové úložiště ve formátu ARC
- 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

Vlastností definice archivního souborového formátu ARC je, že soubor lze identifikovat a extrahovat bez dalších souvisejících indexů, které by tyto soubory popisovaly. Tzn., že pokud není vytvořen příslušný index, v souborech není možné vyhledávat efektivněji než sekvenčním přístupem, což při velkém počtu dat znamená extrémně dlouhý přístup k datům. V takovém případě je velmi náročné zjistit, zda hledaný dokument obsahuje či nikoliv.
Jeden ze způsobů, jak výrazně urychlit přístup k datům, je aplikace nástroje NutchWAX. Tento SW vytvoří fulltextový index, tzn. index, ve kterém lze vyhledávat podle jednotlivých slov. Rámec projektu WebArchiv neumožňuje vytvořit takto rozsáhlý typ indexu pro celý archiv, tzn. sestrojení je výpočetně složité a náročné na fyzické datové úložiště. NutchWAX se tedy využívá pouze na indexování omezené množiny dokumentů (výběrové zdroje se smlouvou a souhlasem od vydavatele k veřejnému zpřístupnění)

Archiv ovšem obsahuje daleko větší množství dat, které nebylo možno efektivním způsobem zpřístupnit. Vznikla tedy potřeba nástroje, který by tuto funkci plnil. Součástí takové aplikace by měl být méně rozsáhlý index, který efektivně vyhledává dokumenty na základě URL a času a tak umožní zprostředkování přístupu ke všem datům. Obrázek popisuje architekturu systému.



A. sklízení elektronických dokumentů z Internetu (**1** - převod dat ze starého formátu nedlib na ARC)
B. fulltextový index NutchWAX, vytvořený pro kolekci nasmlouvaných zdrojů (**3** – předávání kolekce z archivu do NutchWAXe; **4** – přímé předávání dokumentů zahrnutých do výběrových zdrojů z indexu celého archivu)
C. subsystém pro zpřístupnění celého archivu **(2** - aplikace, která vytváří index ze všech archivních souborů ARC)
D. nástroje pro zobrazování dokumentů (**5** - nástroj WERA využívá k dotazování indexu NutchWAX; **6** - modul pro dotazování aplikace Wayback do databáze)

Koncepce a požadavky subsystému pro zpřístupnění celého archivu

Subsystém pro zpřístupňování celého archivu musí splňovat tyto základní požadavky:

1. poskytovat dokumenty z databáze na základě daného URL a času
2. rychlý přístup k datům a odezva
3. schopnost zpracovat rozsáhlé archivy
4. zpracování archivního formátu ARC
5. propojení s aplikačním rozhraním pro zobrazování dokumentů z archivu
6. možnost častého vkládání nových záznamů; inkrementální konstrukce databáze
7. schopnost měnit průběžně vlastnosti záznamů
8. základní správa archivu a operace nad záznamy (statistiky)
9. java web aplikace – multiplatformní
10. modulárnost programu

NÁSTROJE PRO ZPŘÍSTUPNĚNÍ VYUŽÍVANÉ V PROJEKTU WEBARCHIV

Aby byl archiv k něčemu vlastně užitečný, musí být zajištěn přístup k archivovaným dokumentům. Vlastní archivační soubory pouze zajišťují úložiště pro data, ale neumožňují tato data efektivně procházet, nebo v nich dokonce vyhledávat. Pro tyto účely je nutné zkonstruovat index, který se použije pro zpřístupnění, a k němu vytvořit odpovídající rozhraní, které zobrazí výsledky.

Nutch

Nutch patří mezi ambiciózní projekty[13], které se zabývají problematikou vyhledávání na webu. Jde o volně šiřitelné transparentní vyhledávací rozhraní (engine) implementované plně v javě. Systém poskytuje plnou škálu nástrojů, které jsou potřeba k provozování vlastního vyhledávacího enginu. Nutch dokáže:

- stáhnout a zpracovat miliony stránek měsíčně
- spravovat index těchto stránek
- vyhledávat v takovém indexu 1000x za vteřinu
- poskytovat velmi kvalitní výsledky vyhledávání
- minimalizovat náklady na provoz takového systému

Nutch je vhodný pro různorodé prostředí, jednak pro lokální použití v intranetu, taktéž i pro velmi rozsáhlé prostředí celého Internetu. Implementace Nutche vychází z architektury Apache Lucene[14], což je obecné API rozhraní pro indexaci textu a vyhledávání. Další částí Nutche je samotný crawler, který sbírá data a průběžně je indexuje.

NWA

Z iniciativy severských národních knihoven vzešel projekt NWA. Systém se skládal z webového rozhraní napsaného v jazyce php, které umožňovalo zobrazovat archivované dokumenty uložené ve formátu ARC a z java webové aplikace pro zpřístupnění dokumentů ARCRetriever. Index byl sestaven pomocí Apache Lucene. Prototypová aplikace se používala i v projektu Webarchiv, obsahovala velké množství chyb, které bylo třeba odstranit.
Díky spolupráci s norskými vývojáři se jich podařilo spoustu opravit. Zásadní problémy se vyskytovaly v české lokalizaci, správné rozpoznání kódování, vyhledávání českých znaků a zobrazování diakritiky. Do oficiální verze programu byla převzata i kompletní česká lokalizace softwaru. NWA také nesprávně interpretoval javaScript a některé HTML značky (např. rámce). Další nevýhodou, která vedla ke změně indexovacího modulu (spolupráce s IIPC), bylo extrémně dlouhé zpracování indexu při velkém množství dat. NWA projekt byl zařazen do vývojového programu IIPC a části systému byly použity do stávajícího systému WERA.

[13] http://lucene.apache.org/nutch/ ; http://nutch.sourceforge.net/docs/en/about.html
[14] http://lucene.apache.org/

WERA

Spoluprací konsorcia IIPC, Internet Archive a severských zemí vznikla WERA (Web aRchive Access)[15]. WERA je následníkem NWA projektu a využívá jeho hlavní části (prohlížeč archivních dokumentů na způsob Wayback Machine od Internet Archive). Předností tohoto systému je velmi snadná navigace a propracované uživatelské rozhraní (časová osa, která zobrazuje různé časové verze dokumentu ve zvoleném časovém rozlišení).
Výsledky vyhledávání v podobě URL jsou zobrazeny velmi přehledně a u každého odkazu jsou linky na získání dalších časových verzí téhož URL. U odkazu jsou také linky, které vyhledají dané slovo v rámci daného serveru. Zobrazovat archivované stránky lze i pomocí zadání přesné URL adresy. Propojení mezi archivovanými dokumenty a WERA systémem zprostředkovává index Nutche. Kód využívá služeb Nutche, konkrétně XML výstupu servletu opensearchservlet.
Lokalizace češtiny byla částečně převzata z předchozího nástroje NWA a upravena pro aktuální verzi 0.4.2.

Arcretriever

Arcretriever je modul pro dodání dokumentu, jenž je součástí systému WERA. Java web aplikace zobrazuje dokumenty vrácené z archivu. Parametry modulu jsou:

- aid – identifikátor dokumentu tvoří offset záznamu a příslušný název ARC souboru, např. tedy 58657/IAH-2006121416.arc.gz
- reqtype – udává typ vráceného dokumentu. Jednotlivé hodnoty jsou: getfile - vrací původní dokument a getmeta – vrací metadata archivovaného souboru a getfilestatus.

V současnosti se připravuje podpora pro distribuované systémy. Díky tomuto vylepšení nebude modul omezen jen na lokální souborový systém.

NutchWaX

Vývojáři Internet Archive se během roku 2004 rozhodli použít vyhledávací rozhraní Nutch pro potřeby indexování dokumentů archivovaných Heritrixem. Výsledkem tohoto snažení je nástavba Nutche – NutchWaX (NutchWeb Archive Extensions)[16]. Výhodou Nutche je jeho modulárnost, což znamená, že NutchWaX je soubor indexovacích a dotazovacích pluginů, které přidávají do indexu potřebná metadata. Tato speciální pole odpovídají potřebám indexování a dotazování archivních kolekcí dokumentů. Příkladem takových polí jsou jméno archivního souboru, ve kterém je dané URL uloženo, jméno kolekce apod. Dalším podstatným rysem je vlastní zpracování ARC souborů, které jsou pomocí modulu arc2seg konvertovány do nutch-segmentů, které jsou dále indexovány.
V blízké době bude uvolněna nová verze programu, která bude plně distribuovat všechny kroky indexačního procesu a bude podporovat možnost indexovat data z několika různých datových úložišť.

Wayback

Prototypním softwarem, který by měl časem nahradit stávající Wayback Machine, je wayback java aplikace[17]. Aplikace průběžně spravuje index s URL záznamy nad archivními soubory ARC. Spolu s indexačním modulem obsahuje uživatelské rozhraní, které umožňuje zobrazovat dokumenty na základě URL a času a také pomocí tzv. hvězdičkové konvence. Tato konvence dovoluje vyhledávat pomocí levých prefixů URL. Aplikace je v neustálém vývoji a do budoucna se počítá i s možností fulltextového prohledávání.

BUDOUCNOST PROJEKTU

Hlavním cílem bude v roce 2006 spustit a dlouhodobě udržet celoplošnou sklizeň domény .cz. Dále spolu s celoplošnou sklizní nadále sbírat smluvní výběrové zdroje, jejichž význam se stále zvětšuje. Činnost, která představuje proces od výběru zdroje, oslovení vydavatele k podpisu smlouvy

[15] http://archive-access.sourceforge.net/projects/wera/
[16] http://archive-access.sourceforge.net/projects/nutch/
[17] http://archive-access.sourceforge.net/projects/wayback/

s vydavatelem a končí fulltextovým vyhledáváním v archivu, bude třeba zautomatizovat a vytvořit procesní tok událostí.
Pokud bude schválena novela autorského zákona a bude možno legálně nahlížet do kompletního archivu, tedy nejen do nasmlouvaných zdrojů, bude v tomto archivu možno vyhledávat alespoň podle URL a času sklizně dokumentu. Dalšími problémy k vyřešení se ukazují být

- tzv. inkrementální sklízení, což je vlastně problém identifikace změn v opakovaně sklízených dokumentech a následné optimální funkcionalitě crawleru
- duplicita dokumentů. Souvisí s inkrementálním sklízením. Budoucí verze crawleru by měla být schopná rozeznat změny v dokumentech a nesklízet duplicitní dokumenty, na kterých neproběhla žádná změna (např. aktualizace)
- inkrementální indexování. Počet sklízených dokumentů roste a opakované vytváření celého indexu je časově náročné, systém by měl zajistit přidávání nových dokumentů do již existujícího indexu
- identifikace domácích dokumentů mimo doménu .cz. Mimo národní doménu existuje podstatná množina dokumentů s českým obsahem. Řešením může být projít linky ze stažených stránek odkazující mimo doménu .cz a text těchto stránek porovnat nástrojem pro identifikaci jazyka
- fulltextové indexování celého archivu. Vytvořit index pro fulltextové vyhledávání je extrémně náročné na výpočetní čas i na HW. Celý proces je nutno distribuovat na několik strojů. Podporu pro distribuci výpočetních kroků v sobě bude v dalším releasu plně implementovat používaný nástroj NutchWAX
- zakázkové sklízení. Zajištění služby sklízení daných webových portálů a jejich následné zpřístupnění
- analýzy dat. Při převodu ze starého formátu do nového se „velké“ soubory odkládaly do koše. Jednou z analýz bude zjistit, co za data se nachází v těchto koších. Další analýzou může být např. zjišťování reklamních banerů dle velikosti souboru a rozměru obrázku apod.
- otázka trvalých odkazů. Každý dokument v archivu by měl být jednoznačný identifikátor. Pokud budeme na tento dokument odkazovat z nějaké stránky, měl by být tento odkaz perzistentní, tj. neměl by se změnit ani se změnou technologie
- OAI protokol. Archiv se chystá podporovat protokol OAI, který napomáhá komunikaci mezi elektronickými archivy[18]
- v roce 2007 se archiv projektu WebArchiv stane součástí připravované „Digitální knihovny ČR“

[18] http://www.openarchives.org/