Masarykova univerzita
Fakulta informatiky

Integrace a automatizace systémů v pracovních procesech projektu WebArchiv

Diplomová práce

Adam Brokeš

Brno, 2011

Masarykova univerzita
Fakulta informatiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student(ka): Bc. Adam Brokeš Datum: 11. 11. 2010

Program: N-IN Informatika

Obor: Informační systémy

Garant oboru: prof. RNDr. Jaroslav Král, Dr.Sc.

Vedoucí práce: RNDr. Miroslav Bartošek, CSc.

Název práce: Integrace a automatizace systémů v pracovních procesech projektu WebArchiv

Zadání:

Práce by měla podrobně rozpracovat a zdokumentovat probíhající procesy v rámci aktivity WebArchiv Národní knihovny ČR. Posleze identifikovat část procesů, kterou je možno zautomatizovat a tuto část implementovat jako samostatnou jednotku zastřešující ostatní systémy v projektu WebArchiv (WA). Předpokládané výsledky práce:

- detailní popis systémové části projektu WA - workflow, používané nástroje, systémy a procesy,
- popis současného stavu řešení problematiky sklízení webu u předních světových hráčů v dané oblasti (např. Internet Archive, European Archive, významné národní knihovny),
- softwarovy modul, ktery bude automatizovat potřebné činnosti (kopírování a synchronizace souborů mezi sklízecím a zpřístupňovacím subsystémem, indexace, testovací sklizně a případně i sklizně standardní).

Základní literatura:

Souhlas se zadáním (podpis, datum)

student(ka) vedoucí diplomové práce garant oboru

Prohlášení

Prohlašuji, že tato diplomová 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: RNDr. Miroslav Bartošek CSc.

Poděkování

Rád bych poděkoval spolupracovníkům na projektu WebArchiv, kteří se mnou spolupracovali a konzultovali specifikaci systému. Jmenovitě především Ing. Liboru Coufalovi. Dále můj dík patří RNDr. M. Bartoškovi CSc za vedení práce, inspiraci a ochotu.

Klíčová slova

WebArchiv, archivace webu, Grails, webová aplikace, Heritrix, sklízecí robot, Internet, WA Harvester, WA Admin

Shrnutí

Cílem práce je dokumentace a zefektivnění procesů v rámci projektu WebArchiv. Součástí práce je popis stavu současného řešení archivace webového obsahu. V práci jsou rozpracovány přístupy několika světových institucí, které jsou dlouhodobými aktivními hráči v oboru archivace webu. Práce obsahuje aktuální analýzu projektu WebArchiv, která zahrnuje historii a současnost archivu, popis použitých nástrojů, dokumentaci a diagramy jednotlivých probíhajících procesů. V práci je uveden přehled způsobů akvizice dat, spolu se specifiky a dopady na budoucí vývoj. Na základě analýzy procesů je vytvořena specifikace, podle které je navržena aplikace usnadňující správu existujících dat v projektu a zároveň automatizující ručně prováděné úkony. Aplikace je realizována ve webovém frameworku Grails a je rozložena do jednotlivých komponent – modely, řadiče a pohledy – spolu s podpůrnými třídami zajišťujícími další služby, jako napojení na WA Admin a extrakci odkazů z dynamicky generovaných stránek. Cílem práce bylo zároveň vytvoření "manuálu", který by mohl případného zájemce o obor uvést do technických specifik archivace webu v České republice.

Vysázeno v LaTeX

Obsah

Úvod

Internet spolu s informačními technologiemi v posledních dvaceti letech zásadně změnil způsob lidské komunikace i vývoj společnosti jako takové. Toto médium se stalo všeprostupujícím a výskyt různorodých informací na něm obsažených bereme za samozřejmý. Bohužel nelze zaručit trvalou existenci těchto dat. Je zcela běžné, že námi hledané dokumenty již nejsou na původních adresách dostupné, přičemž příčiny mohou být úmyslné i nechtěné. Na straně jedné stojí lidský zásah, kdy se autor nebo jiná osoba rozhodne informace z původního umístění odstranit či přesunout. Existuje ovšem také, a to v nemalé míře, riziko hardwarové chyby, nesprávného zacházení s webovým serverem apod. Z těchto důvodů se nedá spoléhat na trvanlivost dat obsažených na Internetu a pro zachování těchto informací je nutné vyvinout podobné aktivity, jako tomu je u klasických dokumentů (knihy, tiskoviny apod.). V opačném případě může nastat situace, kdy společnost bude hledět na svou nedávnou minulost jako na dobu digitálního temna.

Knihovny zaujímají ve společnosti výjimečné postavení a přispívají k zachování kontinuity lidského kulturního vývoje. Základní rolí knihoven je shromažďování a uchovávání publikací vydaných v dané zemi. Dynamika poslední doby ovšem zásadně mění poměr vydaných tištěných a elektronických dokumentů. Jednoduchost a nenáročnost publikování myšlenek a dokumentů na Internetu vedla k situaci, kdy prakticky kdokoliv s připojením na Internet je sám sobě autorem i vydavatelem. Z těchto důvodů množství publikovaných informací roste závratnou rychlostí a je nutné této změně čelit. Právě knihovny jsou nejvhodnějšími kandidáty pro selekci, akvizici a uchování důležitých dat, která se vyskytují pouze v elektronické podobě, protože neexistuje objektivní autorita, která by nesla zodpovědnost za publikované dokumenty.

Výše zmíněné okolnosti vedly Národní knihovnu České republiky k vytvoření projektu WebArchiv, který vznikl v roce 2000 a od té doby zajišťuje všechny nutné činnosti pro dlouhodobé uchování elektronických informací, které mají kulturní přínos pro český národ. Za tu dobu projekt nashromáždil 35 terabajtů dat, které jsou získávány pomocí softwarového

robota a probíhají v pravidelných kolech (tzv. sklizních). Tyto sklizně kromě samotných stažených dat obsahují řadu metadat, jako například nastavení sklizně, logovací soubory a specifická nastavení pro určité domény. Do současné doby není v projektu zavedena systematická správa těchto informací, a protože jsou z hlediska dlouhodobé archivace důležité, vznikla potřeba aplikace, která by správu těchto dat zajistila a snížila náklady na samotnou akvizici automatizováním vybraných procesů, které je v současnosti nutné provádět ručně. Identifikace těchto dat a procesů v projektu, následná specifikace, návrh a implementace aplikace, která zastřeší vybrané součásti je náplní této práce. Při psaní práce jsem mohl vyžít letitých zkušeností získaných prací na projektu.

První kapitola uvádí čtenáře do světa archivace webových dokumentů. Nejprve nastiňuji vývoj, důvody a způsoby archivace, následuje rozbor významných světových institucí, zabývající se touto problematikou. V druhé kapitole je popsán projekt WebArchiv, jeho historie a současný stav, dále v projektu použité nástroje a probíhající procesy. Specifikace jednotlivých typů sklizní je uvedena ve třetí kapitole. Následuje analýza částí procesů, které lze automatizovat, a jsou vymezeny hranice vyvíjeného systému. Pátá kapitola obsahuje návrh technologického řešení problému, který zvažuje a porovnává existující aplikace, je zde nastíněno uživatelské rozhraní, datový model a architektura. V závěrečné kapitole je popis samotné programové realizace aplikace, která řeší v průběhu práce vytyčené cíle. Tento popis obsahuje výběr programového prostředí a popis samotné implementace služeb a datového modelu.

Kapitola 1

Archivace webových dokumentů

V současné dynamické době je čím dál více jasnější posun od klasických médií k elektronickým. Zatímco před dvaceti lety byly publikace vydávané pouze digitální podobě spíše výjimečné, dnes je to zcela běžný jev a s rozvojem Internetu, elektronických čtecích zařízení, chytrých mobilních telefonů a obecně obrazovek všeho druhu, získává uživatel obrovskou volnost v získávání i konzumaci obsahu. Tato transformace přináší řadu výhod, ale zároveň i určitá rizika. Trh si všímá především výhod a proto je rozmach digitálních medií v poslední době tak živelný, ale vždy by měla existovat nestranná a objektivní instituce, která bude posuzovat dopad těchto transformací na klasickou publikační činnost a společnost jako takovou.

Velké riziko, které se v počátcích digitální éry podcenilo, je životnost těchto digitalizovaných dokumentů. Částečně to bylo způsobeno nezkušeností s novými technologiemi a uvykáním si na změněné poměry. V počátcích informačních technologií většina uživatelů nabyla dojmu, že digitalizací dochází ke konzervaci a z toho důvodu mají elektronické dokumenty delší životnost, než ty klasické. Bohužel je tomu právě naopak. Zatímco klasické dokumenty, které byly napsány před pěti sty lety, jsou dodnes bez obtíží čitelné (pokud byly dodrženy pravidla pro fyzické uchování), elektronické dokumenty, které vznikaly v počátcích Internetu, jsou dnes z velké části již zcela ztraceny nebo jsou uloženy ve formátech, které je dnes možné číst pouze s velkými obtížemi (ať již mluvíme o fyzickém nosiči, či formátu souboru v binární podobě).

Určit dnes přesnou velikost Internetu je z praktických důvodů nemožné. Existuje ale celá řada studií, která se snaží tuto velikost odhadnout. Všechny tyto údaje je nutné brát s rezervou, protože každá použitá metoda má určité nevýhody a jedná se pouze o odhad. V případě veřejného povrchového webu (tedy té částí, kterou lze automatizovaně indexovat) byla v lednu 2005 provedena studie [1], při níž bylo vytvořeno 440 tisíc dotazu v 75 jazycích a výsledkem bylo 11,5 miliardy webových stránek.

Server *WorldWideWebSize.com*[1] počátkem ledna 2011 odhaduje tuto velikost na 13,16 miliard stránek. Na blogu celosvětového vyhledávače Google bylo v červenci 2008 uvedeno [2], že Google zaindexoval 1 bilion jedinečných URL. V tomto množství se ovšem vyskytuje velké množství stránek, které nemají hodnotný charakter (kalendáře, automaticky generované CMS odpovědi).

Internet, jakožto nejdynamičtější médium, je ale velice náchylný na nestálost obsažených informací. Rychlost, s jakou se informace na něm obsažené mění, je natolik vysoká, že existuje jednoznačná potřeba nějakým způsobem tato elektronická data ukládat a archivovat. V opačném případě bychom se za několik desítek let mohli ocitnout tváří v tvář informační krizi, kdy by dnešní doba byla zachycena pouze v zlomku přetrvávajících dokumentů a velká část dat by byla nenávratně ztracena. V následujícím textu se tedy primárně budu zabývat uchováváním obsahu, který je na *World Wide Web* (dále jen webu)[2]. Ten je dnes jednou ze dvou nejvýznamnějších aplikací internetu (druhou je elektronická pošta).

1.1 Historie

Historie archivace webu začíná kolem roku 1996, kdy vznikaly první iniciativy v uchování webového prostoru. Mezi průkopníky tohoto inovativního přístupu patřil *Internet Archive* (IA), Národní knihovna Švédska [3] (Kungliga biblioteket) a Národní knihovna Austrálie (National Library of Australia). Zatímco americký Internet Archive zvolil cestu plošné archivace a idea tohoto archivu je zachování otisku celé lidské společnosti, druhé dva zmíněné projekty mají mnohem užší záběr a zaměřují se především na archivaci zdrojů, které mají pro daný stát kulturní přínos. V roce 1999 s archivací začala Národní knihovna Nového Zélandu a o rok později se připojil např. projekt Národní knihovny České Republiky – *WebArchiv* – nebo také Kongresová knihovna USA. Od roku 2001 počet projektů zabývající se archivací rostl velkou rychlostí a dnes již velká část vyspělých zemí vyvíjí iniciativu v této oblasti. Pro přehled institucí zabývajících se archivací viz příloha A.

Ruku v ruce se vznikem nových projektů zaměřených na archivaci webu šel i vývoj nástrojů k tomu určených a to jak pro akviziční fázi, tak i pro zpřístupnění. V roce 1998, za přispění Evropské komise, vznikl projekt NedLib, který byl koordinován nizozemskou Královskou knihovnou (Koninklijke

1. http://www.worldwidewebsize.com/
2. WWW je pouze jednou z aplikací internetu, ale v souladu se zavedeným běžným používáním je budeme považovat za zaměnitelné.
3. Tento projekt byl v roce 2010 pozastaven z důvodu restrukturalizace knihovny [3]

Bibliotheek) a účastnilo se ho dalších 10 národních knihoven. Tento projekt trval pouze do konce roku 2000 a produktem byl první robot, který byl specializován na archivaci webového obsahu. Systém fungoval na základě analýzy zdrojového kódu stránek a vyhledávání URI v něm obsažených, tímto byl položen základ a vznikla idea *parsovacího* přístupu k akvizici, který v podstatě přetrvává dodnes. Následně od roku 2002 Internet Archive uvolnil první verze svých nástrojů s otevřeným zdrojovým kódem, byly to:

- *Heritrix* – sklízecí robot.
- *Wayback* – nástroj pro zpřístupnění archivu, pracuje nad třírozměrným indexem (je zde doplněn rozměr času).
- *NutchWax* – aplikace pro fulltextovou indexaci dat a následné zpřístupnění indexu (je postaven nad volně šiřitelným vyhledávačem Nutch a také používá třírozměrný index).
- *WERA* (*Web ARchive Access*) – tento nástroj uvádím pouze pro úplnost. Jedná se o dnes již ukončený projekt (vývoj byl ukončen v roce 2006), který nebyl vyvinut IA, ale severskými knihovnami (NWA)[4] a jeho účelem bylo propojit Wayback a NutchWax.

První tři zmíněné nástroje se staly prakticky standardem v institucích, které patří do *International Internet Preservation Consortium* (*IIPC*). Toto uskupení zároveň podporuje a spolufinancuje vývoj těchto aplikací a poskytuje platformu pro výměnu zkušeností a diskuzi o dalším vývoji. S přihlédnutím k významu pro tuto práci provedu popis a analýzu zmíněných nástrojů v samostatné sekci.

1.2 Metody a přístupy k archivaci

Archivace může probíhat na straně *klienta* nebo *serveru*[5]. Na straně serveru existují dvě možnosti, první z nich je, jak uvádí Adrian Brown [5], tzv. *direct transfer*. V tomto případě je přímo ze serveru stažena archivační kopie dat. Druhou možností je *transakční* archivace [6], kdy je na server nainstalován software, který zaznamenává jak požadavky klienta, tak odpověď serveru. Tento přístup by následně dokázal replikovat celou komunikaci, ale nese s sebou několik problémů. Nejprve je nutné, aby správce serveru s archivací souhlasil a umožnil instalaci aplikace, dále archivace zachycuje

4. `http://web.archive.org/web/20030527034926/http://nwa.nb.no/`
5. Celou další kapitolou je archivace databází, viz kapitola 5.6 v [4]

pouze vyžádané stránky, tedy část obsahu může archivaci uniknout a v neposlední řadě je třeba poměrně sofistikovaných nástrojů pro zpřístupnění takto archivovaných dat, protože záznam je nutné rozdělit na podobjekty (požadavek, odpověď) a ty následně zobrazovat podle konkrétního požadavku. Tento přístup se vyskytuje velice zřídka a je uveden pouze pro úplnost, dále se budu věnovat pouze archivaci na straně klienta.

V případě, kdy klient (archiv, instituce) používá pro archivaci software nasazený na vlastní infrastruktuře, hovoříme o archivaci na straně klienta. Při tomto způsobu je spuštěn robot, kterému jsou nadefinovány hranice zájmu a zároveň počáteční URL (*semínko/a*). Ten následně provádí podobný proces jako uživatel, který prochází web, jen s tím rozdílem, že webový *crawler* v zásadě vidí pouze zdrojový kód, což často velice ztěžuje nalezení vnořených URL, jak uvidíme dále. Pro každé URL, které spadá do hranic zájmu, vytvoří HTTP požadavek na cílovém serveru, zachytí doručený obsah a ten zpracuje. Zpracování většinou zahrnuje extrakci vnořených odkazů, které jsou zařazeny do fronty, a permanentní uložení. Tímto způsobem zpracovává celou frontu čekajících odkazů. Při zpřístupnění je nutné přepsat všechny absolutní odkazy, aby navigovaly do archivu a ne na živý web.

Tím jsem odpověděl na otázku *"Jak archivovat?"*, dále se nabízí otázka *"Co archivovat?"*. Přístup většiny archivujících knihoven a institucí zahrnuje následující možnosti, či velice často jejich kombinaci.

1.2.1 Celoplošná archivace

Většinou plně automatizovaný sběr dokumentů, který není vůbec omezen, nebo je vymezen určitou částí Internetu. Sem patří například plošná archivace všech dosažitelných domén, kterou provádí Internet Archive nebo sklizeň celé české domény, tedy těch stránek, jejichž *TLD*[6] má koncovku "cz". V těchto případech dochází k archivaci velkého množství dat. Výhodou jsou relativně nízké náklady na pracovní sílu a zachycení obsahu, který by nebyl vybrán pro výběrové sklizně. Jedná se spíše o otisk stavu webového obsahu (například české domény). Vzhledem k rozšíření složitějších technik a nereálnosti dostatečné kontroly kvality ale jde o otisk poměrně mělký a zprostředkovávající nám spíše představu než přesný obraz. Protože sklizeň trvá v řádech týdnů, nelze ji provádět tak často jako výběrové sklizně.

6. Top Level Domain

1.2.2 Selektivní archivace

Při budování výběrových archivů jsou manuálně (nejčastěji kurátory) cíleně hledány hodnotné zdroje, které splňují selekční kritéria a mají tak předpokládanou hodnotu pro budoucí generace. Tyto zdroje jsou sklízeny s mnohem větším důrazem na přesnost a je zde často prováděna kontrola kvality. Zároveň bývá naplněna potřeba zdroje sklízet častěji, aby byla zachycena evoluce webu a nedocházelo ke ztrátě dat mezi sklizněmi.

1.2.3 Tematické sklizně

Jsou podmnožinou selektivních sklizní, kdy je manuálně (kurátory, uživateli) nebo automatizovaně (pomocí vyhledávání klíčových slov) vytvořen seznam dokumentů a webů, které se týkají určité události, která v dané době získala na významu a existuje zde riziko, že po opadnutí zájmu budou dokumenty nenávratně ztraceny. Od selektivních sklizní se liší tím, že probíhají nárazově a lze je roztřídit do tematických kolekcí. Jako příklad lze uvést české povodně v roce 2002 nebo globální zájem o server WikiLeaks v roce 2010.

1.3 Instituce a projekty

Protože Internet je médium globální a přesahující všechny hranice, nelze ani obsah parcelovat do jednoznačných, ohraničených celků. Z tohoto důvodu vzniklo velké množství institucí, které se archivací zabývají. První impulsy vzešly z řad knihovníků[7], kteří si velice záhy uvědomili potřebu archivace, což je jedním z důvodů, proč velkou část institucí zabývajících se archivací tvoří knihovny nebo je knihovnami podporována. V následující části jsem vybral několik důležitých organizací, které se dají označit za významné hráče na poli archivace webových dokumentů. Seznam webových archivů je možné nalézt v Příloze A.

1.3.1 International Internet Preservation Consortium (IIPC)

V červnu roku 2003 iniciovala *Národní knihovna Francie* (*Bibliothèque nationale de France*[8]) založení konsorcia IIPC[9]. První tři roky bylo členství

7. Brewster Kahle, zakladatel nejznámějšícho webového archivu Internet Archive, je původně také knihovník
8. http://bnf.fr
9. IIPC mělo původně 12 členů; kromě BNF to byly národní knihovny Australie, Kanady, Dánska, Finska, Islandu, Itálie, Norska, Švédska, Britská knihovna, Kongresová knihovna

v organizaci omezeno pouze na zakládající členy, ale postupně došlo k jejímu otevření a dnes mohou o členství požádat knihovny, archivy, muzea a další instituce zabývající se archivací webu jako součásti zachování kulturního dědictví. Organizace sama o sobě neprovádí archivaci, ale je zaměřena především na koordinaci společného úsilí a naplňování vytyčených cílů, kterými jsou:

- Vytvořit podmínky pro vznik kolekce velké časti internetového obsahu z celého světa v takové podobě, aby bylo možné jej archivovat, zabezpečit a zpřístupnit.
- Podporovat a koordinovat vývoj společných nástrojů, technik a standardů, které umožní vznik mezinárodních archivů.
- Celosvětově podporovat a motivovat národní knihovny k řešení problematiky archivace a uchování Internetového obsahu.

Objem práce je rozdělen do tří pracovních skupin pro jednoduchou organizaci a jasné vymezení pracovní náplně.

Harvesting working group

Access working group

Preservation working group

Národní knihovna ČR je členem od roku 2007 a od té doby se aktivně zapojuje do všech tří pracovních skupin.

1.3.2 Internet Archive (IA)

Dnes je *Internet Archive*[10] pravděpodobně největším a nejznámějším internetovým archivem na světě. Jeho počátky leží v roce 1996, kdy Brewster Kahle založil neziskovou organizaci, která si klade za cíl vytvořit digitální archiv všech dostupných informačních zdrojů. Tento úkol je velice obtížný, protože se nejedná pouze o webové dokumenty, ale i knihy, hudbu, filmy, software a každý formát s sebou nese jiné překážky (problémy digitalizační, právní i archivační). V témže roce založil i společnost Alexa Internet, která se primárně zabývala plošnou indexací internetových stránek za účelem archivace webu a poskytování komerčních služeb a sehrála důležitou

(USA) a Internet Archive (USA);http://netpreserve.org
10. http://archive.org

roli v zachování otisku této doby. Používala k tomu vlastního robota, který každé dva měsíce archivoval všechny dostupné stránky (v tehdejší době se jejich počet pohyboval v jiných dimenzích, než je tomu dnes) a index dodávala pro Internet Archive. Později společnost koupil velký internetový obchodní dům Amazon.com. Počátkem roku 2009 IA obsahoval 150 miliard archivovaných webových stránek a objem dat dosahoval 4,5 Petabytů (4,5 milionu Gigabytů)[11].

Od roku 1996 se IA aktivně zabývá vývojem nástrojů, které používá pro archivaci a zpřístupnění webových stránek. Aby dostála své filozofii, vyvíjí tento software v licenci otevřeného zdrojového kódu, což přináší výhody v podobě stability, globálního testování a v neposlední řadě zabezpečení kontinuity vývoje nebo alespoň údržby. Protože se tyto nástroje staly jakýmsi standardem a a jsou používány mnoha institucemi (včetně WebArchivu), budou podrobněji rozebrány v podkapitole Nástroje.

1.3.3 Internet Memory Foundation (IMF)

V roce 2004 byla v Amsterdamu založena nadace *European Archive*[12], která se od té doby aktivně zabývá archivací evropského kulturního dědictví nalézajícího se na Internetu a podporuje tuto aktivitu osvětou a také přímou i nepřímou technickou podporou. Nadace v současnosti archivuje měsíčně několik desítek terabytů webového obsahu a vyvíjí technologie, které by umožňovaly větší objem a kvalitu stažených dat (v tomto směru organizace řeší několik projektů, např. *LiWA* nebo *LAWA* – viz níže). V roce 2010 došlo k přejmenování z European Archive Foundation na *Internet Memory Foundation* (*IMF*), kdy se mírně změnila struktura financování (IMF je financován především na projektové bázi) a zároveň lze zaznamenat posun celkových cílů projektu. Dnes si nadace klade za cíl sloužit jako sdílená platforma pro archivy, vědecké instituce a podobně. Pro tento účel je kritický vývoj škálovatelných aplikací, které se budou schopny vypořádat se stále se rozrůstajícím množstvím webových dokumentů (projekt LAWA). Zpřístupnění archivu je zatím naplánováno na rok 2011 na portálu Archivethe.net[13]. Původně projekt používal robota Heritrix pro akvizici dat a Wayback pro zpřístupňování dat uložených. Dnes, s posunem vize k rozsáhlému archivu, dochází k vývoji vlastních nástrojů, které ovšem ještě nenabraly jasně definované tvary a informace o nich nejsou přístupné veřejnosti.

11. http://www.archive.org/post/243665/wayback-machine-comes-to-life-in-new-home
12. http://europarchive.org
13. http://archivethe.net

Living WebArchives (*LiWA*)

LiWA [14] je tříletý projekt financovaný Evropským společenstvím z prostředků v sedmém vývojovém rámcovém programu. Klade si za cíl umožnit zdokonalení, či celkovou změnu nejpoužívanějšího způsobu archivace webu, kdy jednotlivé sklizně jsou realizovány jako statické snímky dat nacházejících se na serveru. Tento přístup ale s rozvojem Web 2.0 naráží na velké obtíže, protože dnešní webové aplikace jsou často realizovány jako dynamické aplikace[15], které využívají JavaScriptu a obsah nahrávají dynamicky. V té chvíli nedochází ke změně URL a tak je se současnými technologiemi prakticky nemožné takovou stránku archivovat a replikovat její funkcionalitu. Tohoto projektu se účastnila Národní knihovna ČR i Moravská zemská knihovna.

Projekt je rozdělen na šest oblastí a těmi jsou:

- Přesnost – archivace a replikace stránek v maximální možné autenticitě. Klíčové je získávaní dynamických odkazů a simulace AJAX technologie, tedy aspekty, které nelze vyřešit parsovacím přístupem. V projektu vznikne *LinkExtractor*, což je webová služba[16], která po zaslaní URL vrátí seznam URL na ni se nacházejících.

- Spam – modul, který využívá strojové učení a snaží se eliminovat spam v archivu (jak spamové farmy, tak spam obecného charakteru, záleží na počátečním přiřazení rozhodovacího vzorku).

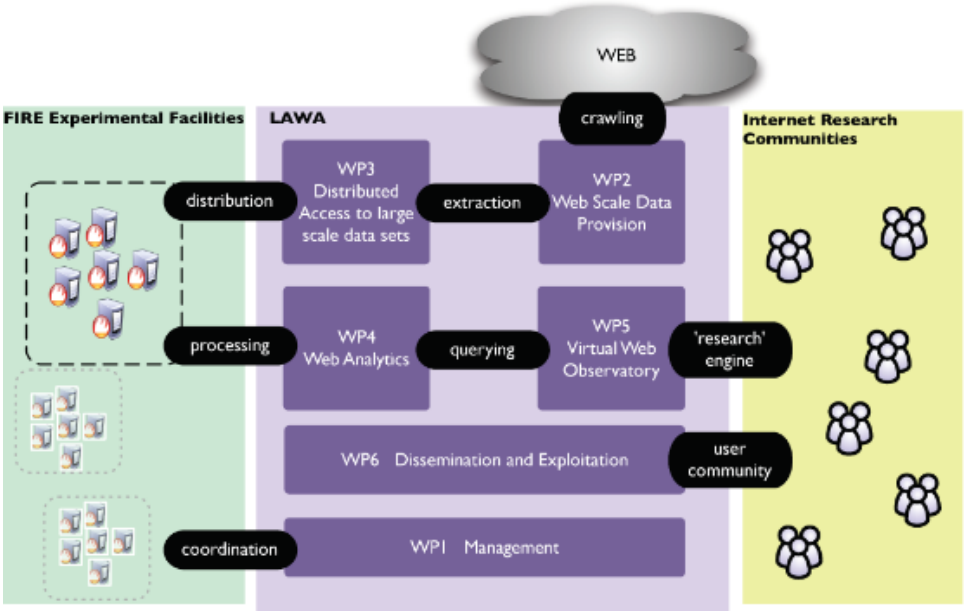
- Časová koherence – při archivaci je nutné počítat s jistou nekonzistencí, protože sklizeň trvá určitou dobu a tak se tedy v okamžiku stahovaní zanořené stránky mohl původní dokument změnit. Temporal Coherence modul řeší situaci tak, že provádí iniciální stažení a zároveň posléze provádí revizi navštívených stránek v opačném pořadí.

- Sémantická evoluce – zkoumá vědecké využití webových archivů k účelu sémantické evoluce termínů (např. “Sankt-Piter-Burh”, “Saint Petersburg”, “Petrograd”, “Leningrad”, “Saint Petersburg”). Je zde využita tematizační metoda a následné hledání vztahů.

14. http://liwa-project.eu
15. Klasickým příkladem může být aplikace *Facebook* (http://facebook.com)
16. Provozována *Hanzo Archives*

- Social web application – tato aplikace se zaměřuje především na zdokonalení archivace sociálních sítí. Tyto sítě umožňují uživatelům tvořit komplexní obsah a velice často používají technologie, které způsobují problémy jak v archivaci, tak zpřístupnění (Javascript, AJAX).
- Komplexní multimediální obsah – řeší problém streamovaných videí a audia, stejně tak jako získání dat od nejrůznějších poskytovatelů komplexního obsahu (youtube.com, video.google.com atd.).

Longitudinal Analytics of Web Archive data (*LAWA*)



Obrázek 1.1: Návrh architektury v projektu LAWA

Tento projekt je především zaměřen na využití archivovaných dat pro výzkumné účely a škálovatelnost rozsáhlých archivů. V projektu bude distribuováno několik jednotek *FIRE*, což je kontejner pro distribuovaný obsah European Archivu. Nad těmito jednotkami bude umožněna analýza dat a za tímto účelem bude rozšířen existující nástroj Hadoop, který bude poskytovat možnost distribuovaného uložení a indexace dat, škálovatelné datové agregace, analýzy dat s přihlédnutím k časovému rozměru a automa-

tické klasifikace webového obsahu. Cílovými uživateli budou především vědci a tyto změny jim umožní provádět cílené analýzy dat nad extrémně velkými bázemi dat, které bude budoucí Internet představovat. V současnosti ale projekt nenabyl zcela konkrétních obrysů a v nejbližší době bude sestaven testovací případ užití, který bude použit k ověření definovaných funkcionalit. [7]

1.3.4 Netarkivet.dk

Od 1. července 2005 platí v Dánsku nový zákon o povinném výtisku, který umožnil vznik projektu *Netarkivet*[17]. Ten je řešen společně Státní a univerzitní knihovnou v Aarhusu a Královskou knihovnou v Kodani. Cílem je provádět sběr a uchování dánské domény a dánských elektronických dokumentů. V současné době zákon neumožňuje data veřejně zpřístupňovat, přístup je umožněn pouze pro výzkumné účely a to s povolením Dánského úřadu na ochranu údajů (Danish Data Protection Agency).

Na projektu pracuje 20 zaměstnanců v rozsahu 4,5 plného úvazku. V polovině srpna 2010 archiv obsahoval přibližně 4,5 miliard dokumentů o objemu 160 TB. Dánská doména obsahuje 1,1 milionu domén druhé úrovně, z nichž je aktivních zhruba milion a k tomuto seznamu je třeba připočítat 45 000 domén, které jsou zaměřeny na dánskou společnost, ale leží mimo doménu .dk. [8]

Archivační strategie:

1. Celoplošná archivace – získání celkového přehledu o doméně .dk, od července 2005 bylo provedeno 9 celkových sklizní. Sklizeň vychází ze seznamu všech domén získaného od administrátora národní domény .dk, k němuž je přidán seznam dalších semínek, zajímavých pro dánskou veřejnost.

2. Selektivní sklizně – cílem je zachytit intenzivně se měnící stránky, tedy místa, kde by nebyly změny zachyceny celoplošnou sklizní. Jedná se o 93 domén, které jsou sklízeny denně, týdně a ročně.

3. Tematické sklizně – archivují informace, týkající se událostí významných pro dánskou společnost, o kterých se předpokládá, že zmizí, až zájem o událost opadne.

17. http://netarchive.dk

V projektu je pro archivaci použit systém *Netarchive Suite*, který byl vyvinut pro účely Netarchive.dk.[18] Z důvodů důležitosti a praktického významu v této práci mu bude věnována samostatná část.

1.3.5 Pandora

Australský projekt[19], který vznikl v roce 1996 původně v Národní knihovně Austrálie (NLA), s cílem vytvoření kolekce kopií australských online publikací (patří tedy mezi průkopníky webové archivace). Dnes se na vytváření kolekce podílí dalších devět dalších australských knihoven a jiných paměťových institucí. Do archivu jsou vybírány, archivovány a následně zpřístupněny dokumenty o Austrálii, psané australským autorem, jejichž téma a obsah je přínosný a relevantní pro Austrálii nebo obecné mezinárodní znalosti. Legislativa Austrálie (konkrétně Autorský zákon z roku 1968) obsahuje položku povinný výtisk, ten se ale nevztahuje na publikace na dostupné z webu. Proto jsou knihovny nuceny uzavírat před akvizicí a zpřístupněním dokumentu smlouvy s vydavateli, podobně jako České republice (viz 2.1.1). Archiv je dnes tvořen selektivními a celoplošnými sklizněmi. Pro komplexní správu selektivních sklizní slouží systém *PANDAS*[20], který popíši v následující podsekci. Celoplošné sklizně jsou prováděny na základě smluvního vztahu se společností Internet Archive. Dosud byly provedeny tři celoplošné sklizně a to v letech 2005, 2006 a 2007. Tyto sklizně nejsou v současné době zpřístupněny z legislativních důvodů a jsou zatím pouze fyzicky uloženy v *petaboxech* [21] v NLA. Statistiky těchto celoplošných sklizní můžeme vidět v tabulce 1.1 a tabulka 1.2 uvádí obsah selektivní části archivu.

Sklizeno	2005	2006	2007
Unikátních dokumentů	185 549 662	529 238 990	516 064 820
Dokumentů celkem	189 824 119	621 664 876	523 510 945
Počet domén	811 523	1 260 553	1 247 614
Nekomprimovaný objem	6,69 TB	19,04 TB	18,47 TB
Komprimovaný objem	4,52 TB	10,48 TB	10,18 TB

Tabulka 1.1: Statistika australských celoplošných sklizní [4]

18. http://netarchive.dk/suite/
19. http://pandora.nla.gov.au
20. http://pandora.nla.gov.au/pandas.html
21. Přenosné diskové pole, které je uzpůsobeno pro uchování velkého množství dat. Blíže viz http://www.archive.org/web/petabox.php

K 28. 11. 2010 archiv obsahoval	Celkem	Měsíční růst
Počet archivovaných zdrojů	26 484	173
Počet archivovaných instancí	59 781	763
Počet souborů	101 890 649	2 052 420
Objem dat	4.57 TB	102 GB

Tabulka 1.2: Statistika australských výběrových sklizní [9]

PANDAS

Tento systém vznikl v roce 2001 a od té doby byl již dvakrát přepsán, poslední verze – PANDAS 3 je z června roku 2007. Účelem systému je zajistit funkcionalitu a automatizaci pracovního procesu, který je využíván NLA a dalšími partnery pro archivaci webového materiálu (aplikace tedy plně podporuje koordinaci vzdálených institucí). Tento systém je v mnohém podobný aplikaci WA Admin, která je použita v projektu WebArchiv. Pro sklízení je použita terminálová verze programu HTTrack[22].

Funkcionalita

- Správa administrativních metadat o zdrojích, které byly vybrány pro archivaci, odmítnutých nebo čekajících na rozhodnutí.
- Správa přístupových omezení k těmto zdrojům.
- Časování a spouštění jednotlivých sklizní.
- Správa kontroly kvality a proces zajištění opravy problému.
- Příprava a organizace archivovaných instancí pro zpřístupnění (podle názvu stánek, předmětového třídění apod.).
- Generování zpráv pro management.

V projektu je zajímavě vyřešena otázka perzistentních identifikátorů, kdy všechny odkazy z archivu splňují definované schéma. [10]

22. http://httrack.com

Kapitola 2

Projekt WebArchiv

2.1 Historie a současnost

WebArchiv [1] vznikl v roce 2000 v rámci programového projektu výzkumu a vývoje Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet pod záštitou Ministerstva kultury ČR, tento grant probíhal v letech 2000 a 2001. Projekt je řešen v Národní knihovně České republiky (NK) a je financován téměř výhradně z grantové podpory Ministerstva kultury ČR. Současný výzkumný projekt Ochrana a trvalé zpřístupnění webových zdrojů jako součásti národního kulturního dědictví započal v roce 2006 a končí v roce 2011. Přímo v NK existuje specializované oddělení archivace webu, které se zabývá pouze touto problematikou. Technická část projektu je realizována na základě konzultací s Moravsku zemsku knihovnou v Brně (MZK), která je jejím koordinátorem. Externím spolupracovníkem je Ústav výpočetní techniky Masarykovy univerzity v Brně (ÚVT), který zajišťuje dodání služeb v oblasti informačních technologií. Na programovém řešení se podílí tým studentů Fakulty informatiky Masarykovy Univerzity.

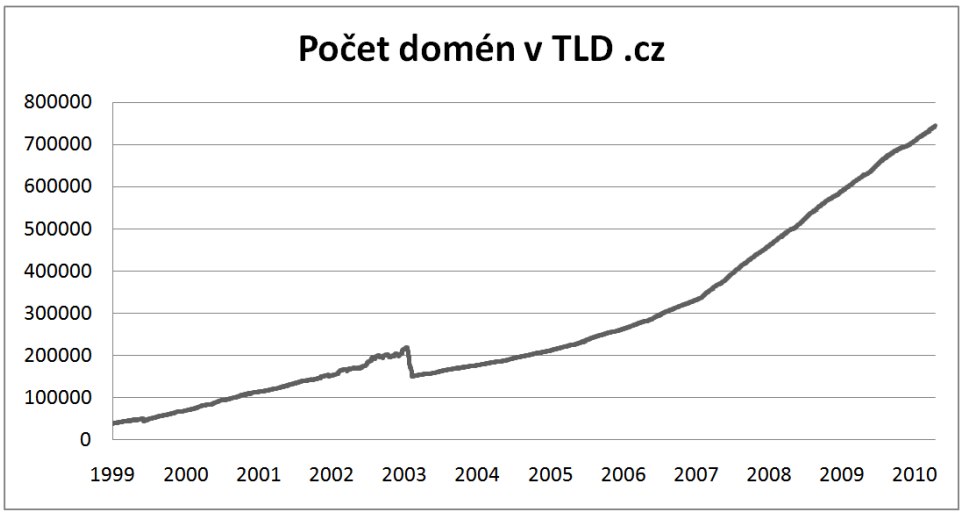
V roce 2000 byl projekt technicky zajištěn jedním serverem umístěným v MZK a páskovým robotem, který se nacházel v Národní knihovně. Sklízení[2] probíhalo nástrojem *NEDLIB Harvester* [11]. Tento robot sloužil dobře pro výběrové sklizně, ale při celoplošném sklizení domény .cz narazil na technické omezení. Robot se po čase zpomalil do té míry, že nebylo možné dále pokračovat ve sklizni. Dnes je již vývoj zastaven. V roce 2004 byl nahrazen programem Heritrix, crawlerem[3] s otevřeným zdrojovým kódem, vyvíjeným pod záštitou Internet Archive.

V roce 2007 Národní knihovna zakoupila datové úložiště pro své projekty a pro WebArchiv vyčlenila na poli 10 TB. Bohužel v témže roce do-

1. `http://www.webarchiv.cz`
2. Sklízení je automatizované shromažďování relevantních elektronických dokumentů
3. Crawler je softwarový robot, jehož primárním účelem je automatizované sklízení elektronických dat

šlo k havárii na úložišti a byla ztracena přibližně jedna třetina dat z celoplošných sklizní. Rok 2008 znamenal pro projekt velký posun, protože došlo k nákupu dvou dedikovaných serverů s kapacitou 20 TB každý. Jeden je umístěn v Praze a slouží především pro testování a indexaci, druhý je umístěn na ÚVT v Brně a jeho primární účel je sklízení a archivování dat. [12] Následujícího roku bylo i na úložišti zpřístupněno 11 TB prostoru a proto bylo možné udělat dosud největší snímek českého Internetu, kdy celoplošná sklizeň dosáhla rozsahu 9,3 TB dat. V roce 2010 došlo k navýšení prostoru na digitálním úložišti o 20 TB a harmonogram sklizní byl ustálen na 6 výběrových sklizní se smlouvou, tři výběrové sklizně bez smlouvy a dvě celoplošné.

Množství českých webových stránek trvale roste, jak je zřetelné například z počtu registrovaných domén (Obrázek 2.1), a je pravděpodobné, že v tomto objemu existuje určitá část dokumentů, které mají přínos a význam pro českou kulturu. Zároveň v kombinaci s celoplošným přístupem archiv zajišťuje pravidelnou archivaci kvalitních a manuálně prověřených stránek, které jsou sklízeny v kratší časové periodě a s mnohem větším důrazem na kvalitu.



Obrázek 2.1: Počet domén v české národní doméně

2.1.1 Legislativní omezení

Strukturu archivu dnes do značné míry formují právní omezení kladená na zpřístupnění archivovaných dat. Podle paragrafu číslo 37 autorského zákona[4] jsou knihovny oprávněny pořizovat archivní kopie a toto ustanovení lze podle některých právních výkladů vztáhnout i na archivaci webu. Avšak k následnému zpřístupnění pro širokou veřejnost již Národní knihovna ČR potřebuje souhlas autora, případně vydavatele, který lze udělit dvěma způsoby. První možností je podepsat s Národní knihovnou tištěnou smlouvu o zpřístupnění obsahu, druhou pak vystavit obsah stránek pod některou z veřejných licencí *Creative Commons* (*CC*) [13]. Obě varianty umožňují zpřístupnit archivovaný obsah na stránkách projektu.

Tento pracovní proces probíhá v praxi tak, že kurátoři vyhledávají zdroje, které splňují selekční kritéria. Následně kontaktují vydavatele vybraného zdroje s nabídkou zařazení zdroje do archivu. Pokud vydavatel souhlasí, může se rozhodnout, zda stránky vydá pod licencí CC nebo podepíše smlouvu. Poté jsou archivovaná data patřící k danému zdroji zpřístupněna.

Přístup k archivním kopiím, pořízeným na základě výše uvedeného paragrafu 37 Aut. zák., je možný pouze z terminálů, které jsou umístěny v NK. Z těchto důvodu je primární zaměření projektu na vybrané zdroje s doplněním celoplošných sklizní.

2.1.2 Statistiky archivu

Celkové statistiky[5]

- Počet souborů: 526 548 170.
- Objem komprimovaných dat: 24 TB.
- Objem nekomprimovaných dat: 35 TB.
- Počet zdrojů se smlouvou: 2333.
- Počet smluv: 1900.

Tematické sbírky

Součástí archivu jsou tematicky zaměřené kolekce, které se týkají v dané době aktuálních událostí. V počátku kurátoři provedou výběr dokumentů,

4. 121/2000 Sb. – Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon)
5. K 17. 12. 2010

které jsou posuzovány a zařazeny podle relevance k tématu. Jedná se tedy o druh selektivních sklizní, které jsou ale zpracovány odlišným způsobem než výběrové sklizně. V Tabulce 2.1 je přehled sklizní provedených do konce roku 2010.

Sklizeň	Začátek	Počet souborů	Objem (MB)
Mistrovství světa v klasickém lyžování Liberec 2009	Leden 2009	97 457	5 121
Volby do Evropského parlamentu 2009	Květen 2009	2 023 399	56 751
České předsednictví EU	Leden 2009	7 230 256	372 845
Výročí obsazení Československa 1968	Srpen 2008	8 941	400
Prezidentské volby 2008	Únor 2008	859 486	18 570
Nová budova Národní technické knihovny	Červenec 2007	5 941	127
Praha olympijská	Duben 2007	8 941	318
Nová budova Národní knihovny ČR	Březen 2007	52 297	1 688
Volby 2006	Květen 2006	1 088 955	54 793
Vysočina	Prosinec 2005	974 912	23 548
Dalimilova kronika	Květen 2005	11 057	472
Povodně 2002	Srpen 2002	2 875	113

Tabulka 2.1: Tematické sbírky

Celoplošné sklizně

Od roku 2001 je pravidelně každý rok provedena jedna celoplošná sklizeň. Výjimkou je rok 2003, kdy sklizeň neproběhla. V počátcích projektu byl nejvíce omezujícím faktorem diskový prostor. Tato situace se od roku 2009 změnila, nejprve v souvislosti s nákupem nových 20 TB diskových polí a posléze vybudováním centrálního datového úložiště Národní knihovny. Proto jsou od roku 2010 naplánovány dvě sklizně ročně.

Sklizeň	Začátek	Počet souborů	Objem (MB)
CZ 2001	Září 2001	3 017 058	106 520
CZ 2002	Duben 2002	10 272 093	315 756
CZ 2004	Březen 2004	32 161 396	1 058 305
CZ 2005	Červen 2005	9 336 123	253 785
CZ 2006	Srpen 2006	70 741 016	3 465 016
CZ 2007	Listopad 2007	81 300 000	3 600 000
CZ 2008	Listopad 2008	78 203 483	3 900 000
CZ 2009	Říjen 2009	205 745 412	9 343 456
Celkem		285 031 169	12 699 382

Tabulka 2.2: Celoplošné sklizně

Výběrové sklizně[6]

Sklizeň	Počet soub.	Objem (GB)
Výběrová sklizeň 02/2009	5 900 660	285 GB
Výběrová sklizeň 04/2009	6 646 960	311 GB
Výběrová sklizeň 06/2009	4 210 470	234 GB
Výběrová sklizeň 08/2009	10 269 291	466 GB
Výběrová sklizeň 10/2009	6 165 763	355 GB
Výběrová sklizeň 12/2009	5 741 645	320 GB
Výběrová sklizeň 03/2009 – bez smlouvy	10 735 277	425 GB
Výběrová sklizeň 07/2009 – bez smlouvy	4 946 054	258 GB
Výběrová sklizeň 11/2009 – bez smlouvy	6 194 482	354 GB
Celkem	282 701 218	12756,9 GB

Tabulka 2.3: Výběrové sklizně 2009

6. Statistiky vznikly analýzou logovacích souborů; u starších sklizní nejsou přesná data dostupná

Sklizeň	Počet soub.	Objem (GB)
Výběrová sklizeň 02/2010	18 719 872	1120 GB
Výběrová sklizeň 04/2010	21 171 938	1311 GB
Výběrová sklizeň 07/2010	13 626 810	523 GB
Výběrová sklizeň 10/2010	18 352 116	1212 GB
Výběrová sklizeň 11/2010	21 816 351	1570 GB
Výběrová sklizeň 03/2010 - bez smlouvy	6 768 333	455 GB
Výběrová sklizeň 08/2010 - bez smlouvy	4 166 805	378 GB
Výběrová sklizeň 10/2010 - bez smlouvy	7 323 428	556 GB
Celkem	111 945 653	7125 GB

Tabulka 2.4: Výběrové sklizně 2010

2.2 Nástroje

Archivace webových dokumentů je činnost technicky velice náročná a existuje množství nástrojů, které jednotlivé úkony provádějí a zjednodušují. V této sekci se zaměřím na nástroje, které se v komunitě, věnující se archivaci webu, staly nepsaným standardem a především jsou použity v projektu WebArchiv.

2.2.1 Heritrix

Heritrix[7] je aplikace pro samotnou akvizici a archivaci internetových stránek a jiného webového obsahu. Vývoj začal v roce 2003, kdy Internet Archive a skupina národních knihoven skandinávských zemí sestavily první verzi dokumentu zachycujícího požadavky kladené na sklízecího robota. O rok později byla vydána verze 0.2.0, která byla první oficiálně veřejnosti přístupnou verzí systému. Systém je soustavně vyvíjen a zlepšován. Vývoj systému se v roce 2009 rozdělil na větev Heritrix 1[8] a Heritrix 3[9]. Verze 2 je dnes uzavřena, protože při vývoji bylo učiněno několik nevhodných návrhových rozhodnutí, ale zdařilé části verze 2 byly integrovány přímo do verze 3. Heritrix je používán v řadě prominentních projektů, jejichž seznam lze nalézt v [14].

Jedná se o aplikaci napsanou v programovacím jazyku Java, což s sebou přináší řadu výhod. Aplikace je multiplatformní a lze ji tedy spustit

7. http://crawler.archive.org/
8. Poslední verze aplikace Heritrix 1 je 1.14.4 – vydaná 10. 5. 2010
9. První a zatím poslední verze je 3.0.0 – vydaná 5. 12. 2009

na různých operačních systémech (Linux, Windows, Mac OS), Java je velice rozšířený, stabilní a populární jazyk[10] a existuje pro něj velké množství knihoven.

Samotný crawler se skládá z modulů, které odpovídají logické sousledností práci s robotem. Nejprve je nutné nadefinovat hranici zájmu (*Scope*), tedy vymezit pomocí pravidel objekty zájmu. Tyto pravidla mohou vypadat následovně:

- Všechny domény, které končí koncovkou .cz;
- obsah podadresáře na adrese http://doména.cz/podadresar/ bez .avi souborů;
- všechny poddomény na adrese http://blog.respekt.cz.

Při vývoji Heritrixu se postupem času (od verze 1.4) upustilo od původních modulů definujících Scope, kterými byly Broad, Host, Domain a PathScope. Tyto moduly omezovaly robota jen na příslušný rozsah (bez omezení, hostitel, doména, cesta) a neposkytovaly dostatečnou flexibilitu pro definici konkrétních hranic, které byly často určeny kurátory při selektivním výběru. Z tohoto důvodu vznikl modul DecidingScope, který o každé URL rozhoduje na základě sady pravidel. Ta jsou uspořádána do řetězce, kdy každé pravidlo může URL označit příznakem REJECT nebo ACCEPT. V konečném efektu je daný dokument stažen či odmítnut na základě finálního příznaku. Tento modul poskytuje široké možnosti nastavení a lze jím simulovat i všechny dříve používané specializované varianty modulu Scope.

Pokud URL patří do oblasti zájmu, pak Heritrix kontaktuje server (většinou použitím http protokolu) a vyžádá si od něj příslušný dokument. Po stažení dokumentu je následně obsah zpracován sadou procesorů typu *Extractor*, které zajišťují extrakci URI odkazů v různých formátech (HTML, CSS, JavaScript, PDF apod.). Všechny získané odkazy jsou zařazeny do fronty a stažený obsah je uložen pomoci modulu *Writer*. Ve verzi 1.14.4 je možné využít ukládání do ARC/WARC souborů, ale také zrcadlit strukturu ve stejné podobě, jako se nachází na serveru (*MirrorWritter*).

Celý proces je spravován modulem *Frontier*, který udržuje seznam stažených a čekajících URL. U těchto adres je prováděna normalizace, například pokud se v adrese změní pouze velikost písmen nebo je zde použit identifikátor *PHPSESSION*, považují se tyto adresy za identické.

---

10. V žebříčku popularity TIOBE je na prvním místě již přes pět let (http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html)

Na tomto místě je vhodné zdůraznit problémy, které s sebou nese použití přístupu, který je založen na strojovém parsování dokumentu. První obtíží je omezená schopnost rozpoznání vnořených odkazů. Pokud je stránka zpracována jednoduchou formou HTML odkazů, které se nacházejí v atributech (*href*, src), pak vše funguje správně. Na problémy narazíme při hledání odkazů, které jsou v dynamicky tvořených a komplexích stránkách. Jako příklad lze uvést elektronický časopis, který zobrazuje jednotlivá čísla v závislosti na rozbalovacím menu, kde uživatel zvolí požadované číslo. Toto dynamické menu Heritrix není schopen simulovat a proto bude sklizena jen úvodní stránka. Technik, které jsou na webu použity a činí Heritrixu problémy je celá řada a jejich popis lze najít v podrobné a kvalitní studii zpracované IIPC [15]. Jedná se o komplexní Javascript, Flash, Silverlight, AJAX a jiné. Heritrix bohužel nepoužívá metodu simulace prohlížeče a proto tyto komplexní technologie není schopen ze zdrojového kódu zpracovat. Tento problém řeší Link Extractor, který je vyvinut v projektu LiWA, viz 1.3.3.

Druhým často se vyskytujícím problémem je sklízení duplicitních dat, která nejsou chybového charakteru, ale neobsahují žádnou relevantní hodnotu. Příkladem může být například dynamický kalendář na blogovém systému. Pokud není zdola ani shora omezen, Heritrix se bude snažit sklízet neomezeně do budoucnosti i minulosti (tedy třeba každý den i po roce 10 000). Je nadmíru pravděpodobné, že v tak vzdálené budoucnosti nejsou zveřejněny žádné příspěvky a tak se bude sklízet pouze informační stránka stejného obsahu ("V tento den nejsou žádné příspěvky"). To může mít za následek zbytečně nadměrné vytěžování stránek a také se může stát, že důležitý obsah bude potlačen na úkor redundantních dat. Oblasti, kde se podobné problémy mohou vyskytovat, jsou kalendáře, přihlašovací stránky, fóra, parametr relace předávaný v rámci URL a jiné (v odborné terminologii jsou tyto problémy souhrně označovány jako *pasti* – ang. Crawler traps).

ARC/WARC formát

Heritrix umožňuje ukládat archivovaná data do formátů *ARC* a *WARC*. ARC formát byl vyvinut v Internet Archive a byl zde použit od roku 1996, do nedávné doby, kdy byl nahrazen formátem WARC. Soubor ARC se skládá z obsahu jednotlivých dokumentů řazených za sebou a každý z těchto záznamů je uveden základními metadaty (url, velikost, čas stažení atd.) a HTTP odpovědí serveru. Takto vzniklý soubor je označen příponou .arc

a je zpravidla komprimován programem gzip[11], z tohoto důvodu soubor má příponu .arc.gz. Velikost souboru je v projektu WebArchiv nastavena na 100 MB, ale v případě vložení extrémně velkého souboru, muže nabýt až několikanásobné velikosti.

```
filedesc://IA-2006062.arc 0.0.0.0 20060622190110 text/plain 76
1 1 InternetArchive
URL IP-address Archive-date Content-type Archive-length

http://foo.edu:80/hello.html 127.10.100.2 19961104142103 text/html
 187HTTP/1.1 200 OK
Date: Thu, 22 Jun 2006 19:01:15 GMT
Server: Apache
Last-Modified: Sat, 10 Jun 2006 22:33:11 GMT
Content-Length: 30
Content-Type: text/html

<html>
Content
</html>
```

Protože metadata obsažená v ARC souborech jsou nedostatečná pro účely dlouhodobého uchování a nejedná se o standardizovaný formát, organizace IIPC v roce 2006 vytvořila návrh nového formátu, který pojmenovala WARC[12] (Web ARChive). Ten obsahuje mnohem bohatší škálu metadat (trvalý identifikátor aj.) a dále umožňuje definici svých vlastních metadat pomocí XML syntaxe ve spojení s definicí *XML Schema*[13] (soubor popisující strukturu XML dat). Tento formát byl v roce 2009 schválen jako ISO standard.

Heritrix 3

Koncem roku 2009 vyšla třetí oficiální verze Heritrixu, který je od základu přepracován v návrhu i použitých technologiích. Ve větvi verze 1 budou pouze odstraňovány chyby, ale nebude docházet již k dalšímu vývoji. Verze 3 zahrnuje následující zásadní změny:

- Je odstraněna podpora JMX a nahrazena architekturou vzdáleného přístupu *REST*.

11. http://www.gzip.org/
12. http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml
13. http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/

- Systém je rozdělen do tří samostatných modulů – Engine, Modules, Common – což zjednodušuje oddělené programátorské zásahy.

- Místo původně speciálního konfiguračního souboru je použit standardizovaný XML soubor, který dodržuje pravidla Spring frameworku (tento soubor má příponu "cxml"). Uložení stavu sklizně, tzv. checkpointing, lze provádět za běhu a bez pozastavení sklizně.

- Došlo ke změně licence na Apache 2 Licence.

- Přidána podpora miliónů vstupních semínek.

- Přidána podpora skriptování – operátor robota se může kdykoliv připojit na konzoli Heritrixu a spustit skript, který bude ovlivňovat chod robota (v současné chvíli je podporováno Groovy, Javascript, Beanshell).

- Přechod webového rozhraní na HTTPS protokol z důvodu bezpečnosti.

- Nový koncept adresáře Action, který umožňuje ovlivňovat chování robota manipulací s tímto adresářem (vkládat nová semínka, vyjmout část stránek ze sklizně).

2.2.2 Wayback

Pro zpřístupnění archivovaných souborů je použita požita javová aplikace *Wayback*, která je nasazena na serveru v javovém kontejneru Tomcat. Tato aplikace, která byla vyvinuta organizací Internet Archive, umožňuje pracovat se soubory ARC a WARC. Vyhledávání probíhá zadáním dotazu na konkrétní URL požadované stránky, na jehož základě jsou v indexu vyhledány všechny časové verze dokumentu. Po zvolení požadované verze se zobrazí obsah, ve kterém jsou dynamicky přepsány všechny URL v dokumentu tak, aby odkazovaly zpět do archivu. Uživatelské rozhraní nyní tvoří časová osa a upozornění, že se návštěvník nachází v archivu. Velký problém v této fázi představuje zobrazování složitých dynamických interakcí (Javascript, AJAX), absolutní odkazy ve Flashi a většina pokročilejších technologií (Silverlight, streamovaná videa apod.). Tento obsah často nelze zobrazit, případně odkazuje na živý web.

2.2.3 ArcWb

Protože původně index Waybacku neobsahoval dostatek metadat o archivovaných souborech a zároveň nebyly přístupné nástroje pro jednoduché vyhledávání v tomto indexu, vznikla potřeba implementovat nástroj, který by tyto nedostatky řešil. Od roku 2006 je proto k indexování archivovaných souborů použit nástroj *ArcWb* vyvinutý Lukášem Matějkou [16]. Indexace probíhá v servletu ArcWb, který běží na *Tomcatu*. Operátor zadá cestu k adresáři, který obsahuje ARC soubory a aplikace sekvenčně rozbaluje jednotlivé soubory a indexuje dokumenty v nich obsažené. O datech jsou zjištěny popisné údaje (mimetype, velikost, hostitel) a ty jsou následně zapsány do *MySQL* databáze ARCRepos. Wayback je napojen na tento index a hledá zadaná kritéria v tabulce docs. Výhodou tohoto přístupu je existence obsáhlé báze metadat o souborech, která může být využita pro statistické účely a dlouhodobé uchování (Long Term Preservation). Nevýhod je bohužel hned několik. S každou novou verzí Waybacku je nutné provést úpravy pro napojení MySQL indexu, samotný index se již v současné době přibližuje limitům MySQL databáze a rychlost vkládání nových souborů je dnes na zlomku rychlosti původní. Dalším problémem je nutnost existence jednoho adresáře, ve kterém se nacházejí symbolické odkazy na všechny zaindexované soubory ARC (časem je možné narazit na limit počtu pevných odkazů v rámci jednoho adresáře linuxové struktury). Aktuálně je rozhodnuto o přechodu na standardní indexovací metodu ve Waybacku a udržování tohoto indexu pouze pro účely dlouhodobého uchování.

2.2.4 NutchWAX

Pro fulltextovou indexaci WebArchivu je použit nástroj *NutchWAX*[14] vyvinutý organizací Internet Archive, který rozšiřuje funkcionalitu internetového vyhledávače Nutch[15] o indexaci (W)ARC souborů a ukládání metadat specifických pro webové archivy.

14. http://archive-access.sourceforge.net/projects/nutchwax/
15. Vydán pod licencí otevřeného zdrojového kódu; http://nutch.apache.org/

pole	popis	příklad
segment	20100326225912	segment, má význam pouze pro nuchwax
title	Národní knihovna	titulek stránky (z obsahu elementu title)
content		textový obsah dokumentu pro generování úryvků
url	http://narodni-knihovna.cz/	URL dokumentu
digest	sha1:NO2WDXITSO6MDWUBNK3BXZAPZCSLQGE6	otisk (hash) z obsahu dokumentu
collection		jméno kolekce (nepovinné)
date	20081018190624	čas sklizně dokumentu
type	text/html	MIME typ dokumentu
length	28138	velikost dokumentu v bytech
boost	5.0	relevance dokumentu pro řazení výsledků, hodnota je rovna log10N, kde N je počet externích odkazů na tento dokument

Tabulka 2.5: Informace obsažené v indexu NutchWAX

Fulltextová indexace se skládá z následujících fází:

1. Import obsahu dokumentů ze souborů (W)ARC – z každého textového dokumentu jsou extrahována metadata, text a v případě HTML stránek ještě navíc odkazy. Výsledky jsou uloženy do tzv. segmentů.

2. Aktualizace databáze crawleru – tato část je nutná pouze při použití crawleru Nutch, ale z jistých technických důvodů ji nelze vynechat.

3. Invertování odkazů – každému dokumentu je přiřazen seznam stránek, které na něj odkazují.

4. Vygenerování pageranku pro hodnocení relevance stránek – je vygenerován textový soubor obsahující na každém řádku: adresa dokumentu, podle které je soubor lexikograficky setříděn a počet externích odkazů (tzn. odkazů z jiných domén), které na daný dokument odkazují.

5. Indexace – ze segmentů, které byly vytvořeny v první fázi, se generuje invertovaný soubor a ke každému dokumentu se navíc ukládá hodnota pageranku. Seznam metadat ukládaných do indexu je v následující tabulce 2.5.

Fulltextová indexace probíhá po částech, výsledné indexy je třeba sloučit do jednoho a následně odstranit z indexu duplicitní dokumenty, které rozpoznáme podle otisku *MD5* [17].

2.2.5 WA Admin

Tento systém byl vyvinut jako nástroj pro správu celého workflow spojeného s akvizicí webových dokumentů v projektu. Částečně přesahuje do oblasti sklízení (z aplikace se generuje seznam semínek pro sklízení výběrových smluv) i zpřístupnění (dynamické ověření smlouvy ve Waybacku). Aplikace je vyvinuta v programovacím jazyku PHP, který usnadňuje okamžitou testovatelnost a pro vývoj webových systémů je velice vhodný. Na začátku vývoje bylo nutné zvolit vhodný aplikační rámec, po několika testech byl zvolen framework KohanaPHP, který má kvalitní objektový návrh a vývojáři usnadňuje práci. Pro okamžitou interakci s uživatelem je využita knihovna jQuery, která logickým způsobem zapouzdřuje nejednotnou implementaci Javascriptu napříč prohlížeči. Datová vrstva je postavena na MySQL serveru, protože je v projektu nejčastěji použitou databází, ale díky datové mezivrstvě frameworku lze aplikaci případně přenést i na jiný databázový stroj. Pro detailnější popis odkazuji na [18], zde se zaměřím pouze na změny a zdokonalení, ke kterým došlo v mezidobí:

- Podpora mazání záznamů v databázi – systém automaticky kontroluje příslušné závislosti a mazání je umožněno pouze uživateli, který má přiřazenu roli "admin". Při mazání je uživatel vždy znovu dotázán a upozorněn na případné dopady, aby nedošlo k nechtěnému mazání.
- Modul pro kontrolu kvality – jedná se o první implementaci prototypu pro usnadnění kontroly kvality. Kurátorům umožňuje při procházení archivu vyplňovat formulář QA pro daný zdroj a zanášet případné problémy. Nevyhovující zdroje se zobrazují v záložce a uživatel s rolí crawl-operator může následně vytvářet testovací sklizně či jinak řešit zjištěné problémy.
- Záložka a tabulka Konspekt – údaje o zdroji byly rozšířeny o položku podkategorie Konspektu, což umožňuje zaznamenávat před-

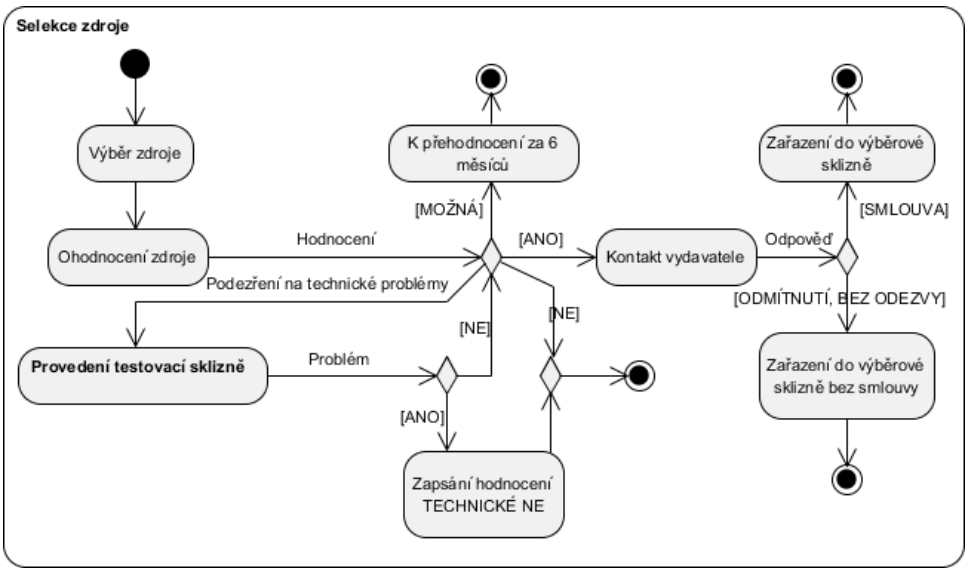
mětové třídění s mnohem vyšší granularitou (24 kategorií X více než 250 podkategorií Konspektu). Na základě tohoto členění vznikla tabulka Konspekt, kde je možné zdroje filtrovat a třídit

- Kterýkoliv kurátor může navrhnout zdroj jako významný a správce příslušné kategorie následně rozhodne o jeho významnosti. Smyslem je nastavení priorit pro akvizici zdrojů tak, aby byly v první řadě podchyceny všechny významné zdroje v dané kategorii.
- Byl zdokonalen hlavní přehled (Dashboard), který nyní zobrazuje všechny otevřené zdroje členěné podle logických kroků workflow se zvláštním upozorněním na akce, které vyžadují pozornost kurátora.
- Zobrazování statistik – nově bylo doplněno generování statistik pro jednotlivé kurátory i celkově s možností dělení podle měsíců a let.
- Vylepšené UI s použitím knihovny JQuery-UI (taby, tlačítka).
- Ve vybraných seznamech zdrojů bylo nově přidáno zobrazení ikon, přehledně indikujících stav zdroje a další důležité informace.

2.3 Pracovní proces

V této části popíši již existující pracovní procesy v projektu, které budou zohledněny při specifikaci systému vyvíjenému v rámci práce. V přiložených diagramech je použit modelovací jazyk UML [19].

2.3.1 Selekce a správa zdrojů



Obrázek 2.2: Diagram selekce zdrojů

V archivu je kladen důraz na kvalitu a vyvážený výběr zdrojů, které se pravidelně archivují. Je to dáno jednak legislativním rámcem, který jsem popsal výše a také rozsahem českého internetového obsahu. Za daných podmínek je nezbytné pečlivě vybrat takové tituly, u kterých se kontaktování vydavatele a důkladná archivace vyplatí a bude z hlediska informačního obsahu přínosem.

Pro vyvážené doplňování zdrojů ve všech oborech lidského vědění je používáno mezinárodní předmětové třídění Konspekt[16]. Každý kurátor má přiděleny určité kategorie Konspektu (tedy např. Kurátor A má na starosti

16. http://konspekt.nkp.cz/

Přírodní vědy a Filozofii) a jeho úkolem je systematicky vyhledávat významné webové zdroje, které se vztahují k těmto tématům/oborům. Nalezené zdroje jsou vloženy do systému WA Admin, kde jsou následně ohodnoceny všemi kurátory.

Z průměru všech hodnocení vznikne výsledné ohodnocení, tedy schválení nebo odmítnutí nominace zdroje. Je-li výsledné hodnocení nerozhodné, zdroj se uloží s příznakem, že je nutné ho za určitý čas (obvykle šest měsíců) přehodnotit. Pokud se v průběhu hodnocení objeví podezření na možné technické problémy, které by se mohly vyskytnout ve fázi zpřístupnění nebo sklízení, požádá kurátor o provedení testovací sklizně, která je zaindexována v testovacím Waybacku. Na jejím základě se pak rozhodne, zdali jsme schopni zdroj archivovat. V opačném případě se uloží hodnocení zdroje "Technické odmítnutí".

Pokud je nominace zdroje schválena, kontaktuje kurátor, který má na starost danou kategorii, vydavatele prostřednictvím emailu, ve kterém popíše projekt a požádá o souhlas se zařazení zdroje do archivu. Souhlas může být udělen podepsáním smlouvy, kterou si vydavatel stáhne ze stránek projektu a zašle do Národní knihovny, nebo vystavením obsahu pod některou z veřejných licencí Creative Commons. Pokud vydavatel udělí souhlas, je zdroj zařazen do pravidelných výběrových sklizní a je možné archivované dokumenty z tohoto zdroje bez omezení zpřístupňovat. V opačném případě zdroj spadá do výběrových sklizní bez smlouvy a dokumenty jsou přístupné pouze z terminálu NK. Celý proces je znázorněn na diagramu 2.2.

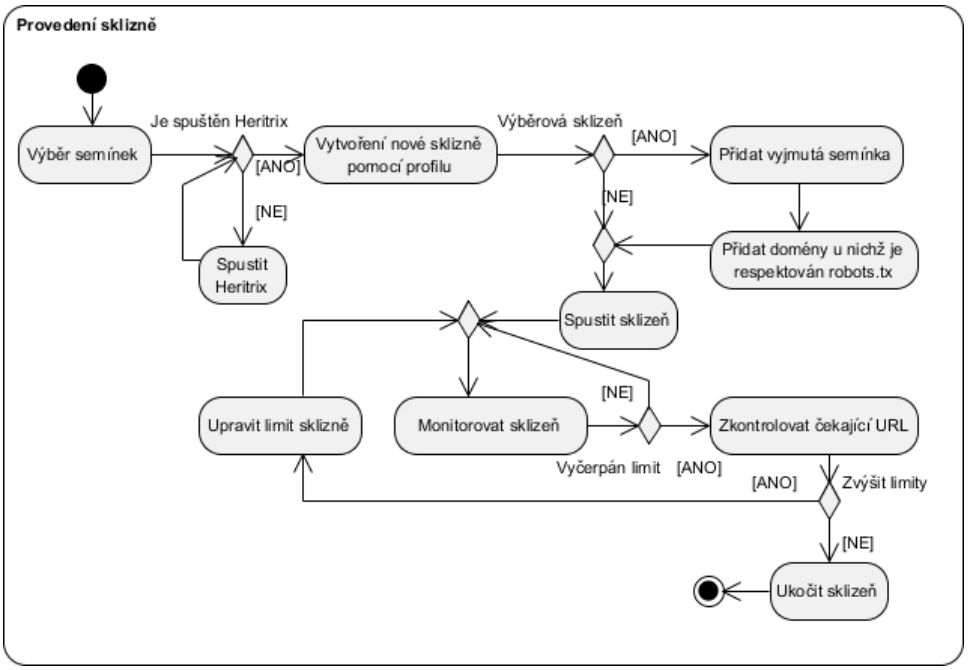
- Použité nástroje: *WA Admin*
- Aktéři: kurátoři v NK

2.3.2 Sklízení zdrojů

Při samotné archivaci zdrojů je využíván Heritrix verze 1. Podle prováděného typu sklizně (více v Kapitole 3) operátor robota získá požadovaná semínka. Z příslušného profilu vytvoří novou sklizeň a vloží získaná semínka. Následně sklizeň monitoruje a opravuje případné problémy. Každá sklizeň má nastaven limit stažených dokumentů pro jednu doménu (tzv. total budget), při jehož vyčerpání se sklizeň pozastaví a operátor může manipulovat s frontami, limit zvýšit, či případně sklizeň ukončit. Tato část není systematicky řízena a do budoucna je nutné vytvořit možnost správy jak sklizní, tak profilů. Podrobněji je proces znázorněn v diagramu 2.3.

- Použité nástroje: *Heritrix 1*

- Aktéři: kurátoři s technickým školením, operátor sklizně



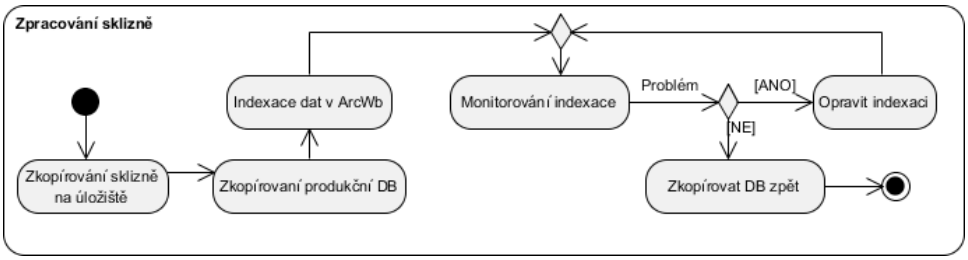
Obrázek 2.3: Diagram sklízení zdrojů

2.3.3 Zpracování sklizně

Po ukončení sklizně operátor přesune data na úložiště, což je digitální archiv zřízený NK za účelem uchování digitálních dat . K tomu požívá běžných linuxových nástrojů, především scp. Následně je nutné data zaindexovat do databáze ARCRepos, která je základní bází dokumentů pro Wayback. Indexace probíhá pomocí nástroje ArcWb a bohužel je při ní zablokována databáze i pro čtení. Protože indexace trvá v řádech několika dnů až týdnů, nebylo by možné v této době poskytovat přístup do archivu. Situace je řešena zkopírováním produkční databáze na server, kde je plánována indexace. Operátor spustí indexaci nad duplicitní databází, zároveň tuto indexaci periodicky kontroluje a řeší případné chyby. Po kompletní indexaci je databáze opět zkopírována na produkční stroj, čímž jsou nové sklizně zpřístupněny pro uživatele. Tento proces je zachycen na Obrázku 2.4 Wayback je propojen s modulem ContractManager, který dynamicky rozhoduje,

zda je možné hledanou stránku zobrazit online nebo pouze na terminálech v knihovně. Toto vyhodnocení probíhá s využitím databáze WA Admin, kdy je kontrolováno, jestli pro danou doménu existuje smlouva (resp. souhlas vydavatele).

- Použité nástroje: *SSH*, *ArcWb*, *Wayback*, *AutoContractManager*
- Aktéři: operátor sklizně



Obrázek 2.4: Diagram zpracování sklizně

2.4 Manuální operace

V následující části jsem sestavil seznam činností, které jsou prováděny manuálně a lze tedy zvažovat jejich automatizaci. Automatizované části budou definovány v rámci specifikace vyvíjeného software.

Nastavení Heritrixu

- Spuštění Heritrixu přes linuxovou konzoli – provádí se jednorázově (10 min).
- Získaní a příprava semínek pro celoplošnou sklizeň je provedena pomocí sed, awt a jiných linuxových nástrojů (1 den).
- Příprava semínek pro výběrové sklizně – použití skriptu (20 minut).
- Vytvoření profilu pro sklízení zahrnuje několik dílčích úkonů, jedná se o zapojení – zřetězení – modulů, které definují práci sklízecího robota (několik hodin pro výběrové sklizně).

- U celoplošné sklizně je nejnáročnější nastavení a vybalancovaní robota, jedná se o sérii experimentů a testů, kdy se ověřuje vliv jednotlivých nastavení na průběh sklizně. Díky již nabytému know-how je tato část o něco zjednodušena, ale vždy je nutné nastavení aktualizovat pro novou verzi robota a zároveň optimalizovat Heritrix od poslední sklizně – (cca 1 týden).
- Při sklízení domény, která vyžaduje speciální nastavení, se vytváří zvláštní konfigurační soubor, který se nazýva override (platný pouze pro danou doménu), ve kterém lze jednotlivá nastavení dále upravovat a využívat dědičnosti – provádí se pro nová semínka každou sklizeň (iterativně 20 a více minut – při důkladné sklizni je tato činnost po celou dobu monitorování robota, tedy několik dní).
  - Vyhnutí se pastím u domény – je nutná znalost regulárních výrazů a jejich využití v jazyku Java (15 minut na past – v případě rozpoznání pasti operátor musí vytvořit specializované nastavení pro doménu a tím je prodloužena doba sklizně).
- V průběhu sklizně je třeba robota monitorovat a řešit případné problémy,
  - pro celoplošné sklizně je třeba monitorovat průběh sklizně po dobu cca jednoho měsíce a tato činnost zabere přibližně týden čistého času; řeší se převážně jen problémy s robotem a nastavením,
  - pro výběrové sklizně tato činnost trvá několik dní, kdy se řeší problémy s jednotlivými doménami (pomocí selektivních nastavení, viz výše).

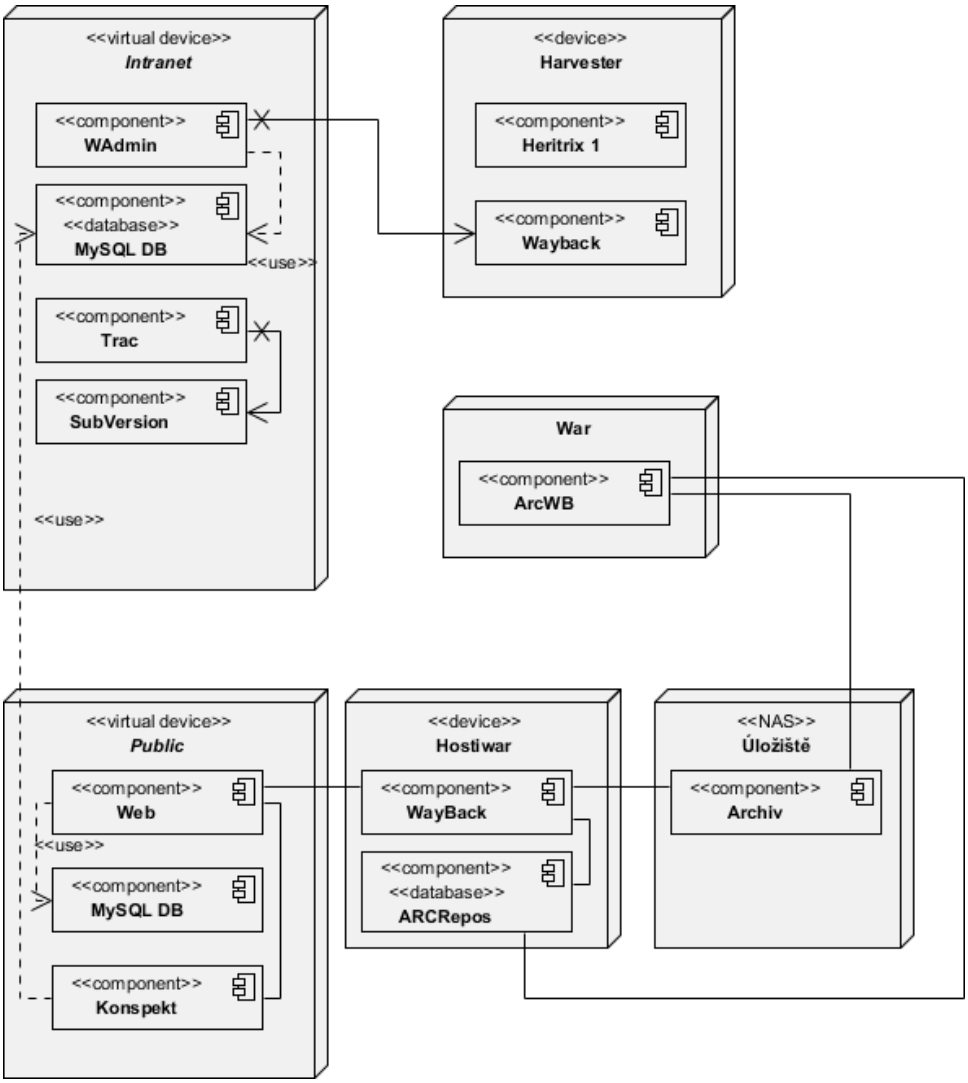
PostProcessing

- Analýza logů po sklizni – je nutné identifikovat strukturu URL a najít vzor případných problémových URL (zde pomáhá přehled o dostupných standardních CMS a různých webových frameworků), prozkoumat návratové kódy HTTP protokolu a revidovat cesty, které vedly k nalezení stránek (transcluded a speculated links) – provádí se každou sklizeň (10+ minut na problematickou doménu, při důkladném šetření cca několik dní).
- Kontrola kvality (QA) – analýza jednotlivých stránek pro QA vyžaduje průzkum DOM modelu (s použitím pomocných rozšíření,

např.: Firebug), identifikaci použitých technologií a do jisté míry závisí na zkušenostech pracovníka, protože občas je nutné vytvořit prvotní odhad toho, kde je problém a podle toho následně Heritrix překonfigurovat (30+ min, při důkladně kontrole celé jedné výběrové sklizni cca 1 - 2 měsíce - viz níže).

- Přesun dat na úložiště – přesuny sklizní jsou prováděny v terminálu a je nutné vytvářet MD5 kontrolní součty pro vyšší bezpečnost (4 hodiny + monitoring; v případě celoplošných sklizní kvůli objemu dat cca 2 dny práce v průběhu 1 týdne).
- Indexace se v současné chvíli provádí přes aplikaci ArcWb, která není zcela stabilní a proto je nutné kontejner (Tomcat) pravidelně monitorovat. V podstatě prochází zadaný adresář a indexuje do zadané kolekce, zde je především důležité zachování struktury úložiště při vkládání nových sklizní (4 hodiny + monitoring; v případě celoplošných sklizni cca 4 dny práce v průběhu cca 2–4 týdnů).
- Fulltextová indexace
  - Je použit systém NutchWAX a v současné chvíli se jedná jen o experimentální provoz.
  - Indexace výběrových sklizní, cca 1500 ARC souborů za jeden den. Výsledný index se pak musí sloučit s dosavadním, to zabere maximálně jeden den. Při indexaci je lepší rozdělit ARC soubory do dávek (optimální množství je cca 2000 ARC souborů) a výsledky pak sloučit dohromady. Není to ale nutné. Při indexaci je třeba dávat pozor na špatné ARC soubory, NutchWAX se s tím není schopen vyrovnat a spadne. Proces je automatický, NutchWAXu se předá seznam ARC souboru a on je zaindexuje. Odhadem cca 1 hodina práce na sklizeň.

2.5 Diagram nasazených nástrojů



Obrázek 2.5: Deployment diagram - Tento diagram popisuje současné rozmístění interních a veřejných služeb.

Kapitola 3

Profily sklizní v projektu WebArchiv

Protože se mnou vyvíjený nástroj bude primárně věnovat provádění určitých typů sklizní, je vhodné na tomto místě podrobněji popsat jednotlivé druhy. Každý typ má uvedeny tyto údaje:

- Základní charakteristiku;
- způsob získání semínek;
- frekvenci sklízení;
- způsob zpřístupnění;
- případné použití deduplikace;
- další případné omezení.

3.1 Výběrové sklizně

- Pravidelná sklizeň zdrojů, na které má NK podepsána smlouvu a může je proto plně zpřístupňovat.
- Počáteční semínka jsou vybrána z báze WA Admin.

```
SELECT s.url
FROM seeds s, resources r
WHERE r.id = s.resource_id
 AND r.resource_status_id = 5 -- Zdroje schválené kurátory
 AND r.contract_id IS NOT NULL
 AND seed_status_id = 1 -- Semínka s~příznakem include
 AND (
  valid_to > CURDATE( )
  OR valid_to IS NULL
 )
 AND (
  valid_from < CURDATE( )
  OR valid_from IS NULL
 )
```

- V současnosti sklizně probíhají každý druhý měsíc.

- Od roku 2011 je naplánovaná sklizeň na každý měsíc s těmito podmínkami[1]
  - Každý zdroj má ve WA Admin přiřazenu frekvenci, tato frekvence může nabývat hodnot - [1, 2, 6, 12 měsíců, jednorázově];
  - v daném měsíci jsou do sklizně zařazeny jen ty zdroje, které byly naposledy sklizeny před dobou, která je větší nebo rovná přiřazené frekvenci sklízení.
- Tyto zdroje jsou zpřístupněny v plném rozsahu.
- Při sklizních je použit modul DeDuplikator a po ukončení sklízení je novými daty aktualizován existující index selected používaný tímto modulem. Výjimkou je sklizeň po půl roce, kdy je modul vypnut.
- Omezení:
  - Nastavení limitu na 25 000 objektů na doménu.
  - Robot má pomocí regulárních výrazů nadefinovány zakázané oblasti. Tyto oblasti mají v databázi WA Admin nastaven příznak exclude.
  - U zdrojů, které mají ve WA Admin nastaven příznak robots.txt, tzn. s vydavatelem bylo dohodnuto dodržování těchto pravidel, je nutné nastavit akceptaci pravidel ze souboru robots.txt uloženého na dané doméně, v ostatních případech se robots.txt ignorují.

3.2 Výběrové sklizně bez smlouvy

- Pravidelná sklizeň zdrojů, které byly vybrány a schváleny kurátory, ale vydavatel smlouvu o zpřístupnění nepodepsal (bud'to přímo odmítl nebo, častěji, oslovení bylo bez odezvy).
- Počáteční semínka jsou vybrána z báze WA Admin.

```
SELECT s.url
FROM seeds s, resources r
WHERE r.id = s.resource_id
 AND r.contract_id IS NULL
 AND rating_result = 2 -- Zdroje hodnocené jako ANO
 AND seed_status_id = 1 -- Semínka s~příznakem include
```

1. SQL dotaz pro výběr semínek bude upraven odpovídajícím způsobem.

```
AND (
  valid_to > CURDATE( )
  OR valid_to IS NULL
)
AND (
  valid_from < CURDATE( )
  OR valid_from IS NULL
)
```

- Zdroje se sklízejí v intervalu 4 měsíce (březen, červenec, listopad).
- Tyto zdroje jsou zpřístupněny pouze na terminálech v NK.
- Při sklizních je použit modul DeDuplikator a po ukončení sklízení je novými daty aktualizován existující index selectedNoContract používaný tímto modulem. Výjimkou je první sklizeň v novém roce, kdy je modul vypnut.
- Omezení
  - Nastavení limitu na 25 000 objektů na doménu.
  - Robot má pomocí regulárních výrazů nadefinovány zakázané oblasti. Tyto oblasti mají v databázi WA Admin nastaven příznak exclude.

3.3 Celoplošná sklizeň

- Pravidelná sklizeň všech webových stránek, které se nacházejí na TLD .cz. Smyslem celoplošných sklizní je podchytit co největší množství "bohemikálních"dokumentů[2].
- Seznam všech českých domén je získán od sdružení CZ NIC[3].
- Sklizeň je prováděna dvakrát ročně, v intervalu půl roku.
- Domény, na které je podepsána smlouva, jsou zpřístupněny v plném rozsahu. V ostatních případech je přístup možný pouze na terminálech NK.

2. Zdroj je definován jako “bohemikální” pokud splňuje alespoň jedno z těchto kritérií: autor obsahu zdroje pochází z České republiky, vydavatel má své sídlo na území České republiky, zdroj je v českém jazyce, zdroj obsahuje významné informace o České republice nebo o českém národu [20].
3. Seznam je získáván ve formátu DNS tabulky bezplatně, na základě smlouvy s CZ NIC. Tento soubor obsahuje množství nadbytečných dat.

- Sklizeň nepoužívá deduplikaci
- Tato sklizeň je prostorově i časově velice náročná, proto je v současnosti zvolen limit 5 000 stažených objektů na doménu a zároveň jsou akceptována pravidla v robots.txt. Je pravděpodobné, že s rozvojem infrastruktury projektu dojde k navýšení tohoto limitu.

3.4 Celoplošná sklizeň mimo doménu .cz

- Výše zmíněná celoplošná sklizeň národní TLD .cz nepokrývá bohemikální dokumenty nalézající se na jiných, generických TLD (např. .eu, .org, .int). Tento "triviální"způsob identifikace žádoucího obsahu pomocí lokace na národní TLD používá z praktických důvodů naprostá většina institucí, provádějících celoplošné sklizně (některé doplňují seznamy semínek o ručně vybrané domény, nacházející se mimo TLD). Zcela automatizovanou sklizeň dokumentů, splňujících výběrová kritéria a zároveň se nacházejících vně národní domény, se zatím systematicky nepodařilo vyřešit žádné instituci. Národní knihovna ČR, jako průkopník v této oblasti, pro tento účel vyvinula prototyp nástroje WebAnalyzer, který umožňuje automatickou analýzu obsahu elektronických dokumentů a určit ty, které jsou zajímavé z hlediska (libovolně) zadaných parametrů. Jedním z hlavních využití tohoto nástroje je identifikace zdrojů, které mají bohemikální či jiný národní charakter. Modul je zatím ve fázi intenzivního experimentování, WA Harvester tedy zatím nelze uzpůsobit přímo pro potřeby tohoto modulu, ale měl by umožňovat případnou integraci tohoto druhu sklizní v budoucnu.
- Semínka se získávají z neakceptovaných URL v celoplošné sklizni; zároveň se provádí experiment s pouze několika semínky význačných webů (seznam.cz).
- Sklizeň je prováděna jednou ročně.
- Domény, na které je podepsána smlouva, jsou zpřístupněny v plném rozsahu. V ostatních případech je přístup možný pouze na terminálech NK.
- Sklizeň nepoužívá deduplikaci.
- Sklizeň zatím nemá přesně známé limity. Z realizovaných testovacích sklizní je zřejmé, že analýza stránek bude paměťově i výkonově

velice náročná.

3.5 Archivační sklizeň

- Jednorázová a mimořádná sklizeň, která je vyžádána ze strany vydavatele nebo kurátorů ve speciálních případech (stránka bude rapidně změněna, přestane být vydávána atd.).
- Semínka jsou v řádu jednotek, většinou pouze jedno a je získáno od kurátora nebo vydavatele.
- Domény, na které je podepsána smlouva, jsou zpřístupněny v plném rozsahu. V ostatních případech je přístup možný pouze na terminálech NK.
- Sklizeň nepoužívá deduplikaci a případné limity vyplývají z konkrétního případu.

3.6 QA sklizeň

- Sklizeň, která je provedena pro zdroje, u kterých se objeví při pravidelné sklizni nějaký závažnější problém, který je třeba testovat novou sklizní. Těchto sklizní může být pro jeden zdroj několik, dokud není problém odstraněn. Posléze je správné řešení přidáno do profilu SelectedSites.
- Semínka jsou v řádu jednotek, většinou pouze jedno a je získáno z aplikace WA Admin, případně z ticketu v systému Trac .
- Frekvence není předem definována, QA sklizně se dělají podle potřeby pro vybrané problematické zdroje po každé sklizni seriálů se smlouvou.
- Výsledky sklizně jsou přístupné pouze v interním WayBacku.
- Sklizeň nepoužívá deduplikaci a případné limity vyplývají z konkrétního případu.

3.7 Testovací sklizeň

- Mimořádná a jednorázová sklizeň vyžádaná kurátory při hodnocení zdroje, který používá sofistikované nebo nestandardní technologie a

postupy, pro otestování, zda je robot schopen jej správně posklízet a WayBack zpřístupnit; provádí se před zařazením zdroje do pravidelných sklizní.

- Semínka jsou v řádu jednotek, většinou pouze jedno a je získáno od kurátora, případně z ticketu v systému Trac.
- Řádově probíhá několik sklizní týdně.
- Výsledky sklizně jsou přístupné pouze v interním WayBacku.
- Sklizeň nepoužívá deduplikaci a případné limity vyplývají z konkrétního případu.

Kapitola 4

Specifikace systému WA Harvester

Ze získaných informací o pracovním procesu v projektu je možné definovat hranice automatizace a data, která lze spravovat. Za tímto účelem jsem v rámci práce specifikoval a navrhl aplikaci, která slouží jako mezivrstva vložená mezi WA Admin a nástroje pro sklízení. Systém má za úkol automatizovat a zjednodušit rutinní práce operátora sklízecího robota. Ten v současné době musí vykonávat řadu úkonů, které lze provádět automaticky. Díky ušetřenému času bude mít možnost se věnovat více kreativní a kvalifikovanější prácí, jako je kontrola kvality sklizených stránek a zdokonalování sklízecích profilů. Použití aplikace by nemělo být příliš složité a i kurátor s průměrným technickým základem by měl mít možnost obsluhovat základní funkce systému. Systém byl pojmenován *WA Harvester* a v následujícím textu bude označován zkratkou *WAH*[1].

Automatická správa má i další přínos, a to v oblasti systematického uchovávání informací o sklizních. Doposud se jednotlivé adresáře, logy, nastavení provedených sklizní uchovávají nesystematicky a tak může v budoucnosti dojít k tomu, že archiv bude obsahovat ARC soubory, ale nebude již dostupná informace o tom, jak byla tato archivovaná data získána. Při používání jednotného systematického uchování informací o sklizních v databázi se nebezpečí ztráty dat minimalizuje.

4.1 Sběr požadavků

Základem návrhu každé aplikace musí být definice účelu, pro který daný software vzniká. Idea a účel WA Harvesteru jsou popsány v úvodu kapitoly, dále rozpracují funkční požadavky, tedy funkce, které bude software zastávat a vykonávat. Při definici funkčních požadavků jsem vycházel z analýzy pracovního procesu v projektu, který je popsán v podkapitole. Zde jsem identifikoval procesy, které je nutné provádět manuálně a které by bylo

1. Název vznikl kombinací WA jako WebArchiv a Harvester jakožto označení robota určeného pro sklízení.

možno automatizovat a tím zvýšit nákladovou efektivnost.

4.1.1 Funkční

Systém bude umožňovat tyto funkce:

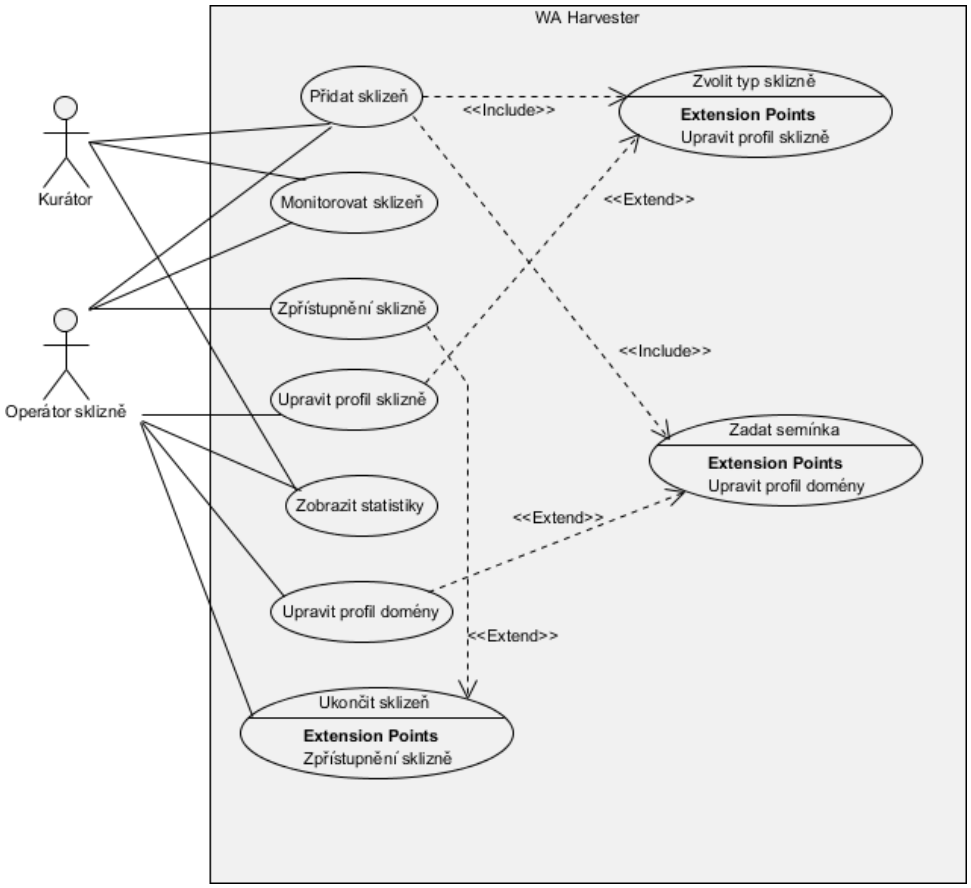
1. Nastavení a spuštění sklizní
   - výběrové se smlouvou
   - výběrové bez smlouvy
   - celoplošné
   - jednorázové archivace
   - QA
   - testovací
2. Monitorování sklizní
3. Ukončení sklizní
4. Zobrazení informací o provedených sklizních
5. Správa profilů pro sklizně
   - úprava obecných profilů Heritrixu
   - úprava profilu pro konkrétní doménu
6. Tvorba statistik (sklizně, archiv)

4.1.2 Automatizace pracovního procesu WA

Systém by měl zabezpečit zautomatizování některých činností. Výhodou automatizace není pouze úspora nákladů, ale také systematické zpracování dat, které snižuje riziko lidské chyby. To je, když vezmeme v potaz dlouhodobost projektu, velice vysoké (ať již v důsledku fluktuace pracovníků, či změny pracovních postupů). Těmito činnostmi jsou:

- Získání semínek (případné formátování, např. CZ doména),
- spuštění indexace sklizní (např. skript, Java třída),
- přesun sklizní na archivační úložiště,
- kontrola integrity přenesených dat.

4.2 Případy užití



Obrázek 4.1: Případy užití systému WAH

Kapitola 5

Návrh systému WA Harvester

Při návrhu systému bylo nutno vycházet z reálného prostředí již nasazených aplikací (viz. Obrázek 2.5) a zároveň prozkoumat trh již existujících řešení. Po důkladném výběru se okruh řešení zúžil na dvě možnosti. Prvním z nich je napojení dánského systému *NetarchiveSuite* na *WA Admin* a na druhé straně stojí vlastní aplikace, která bude komunikovat jak se systémem WA Admin, tak s poslední, třetí, verzí Heritrixu. Ve finále jsem se z praktických důvodů rozhodl pro druhou variantu. V následujících dvou kapitolách popíši výhody i nevýhody zmíněných řešení.

5.1 NetarchiveSuite (NS)

NetarchiveSuite[1] je systém původně vyvinutý dvěma dánskými knihovnami (Státní a univerzitní knihovnou v Aarhusu a Královskou knihovnou v Kodani) pro účely správy sklizní. Je velice robustní a umožňuje celou škálu funkcí. Tyto funkce jsou rozděleny do jednotlivých modulů, které jsou znázorněny v následujícím diagramu. V poslední době se k vývoji nástroje NetarchiveSuite připojily další dvě instituce, národní knihovny Francie a Rakouska. Z tohoto důvodu se vývoj posunul i ve věcech jako je zpřístupnění a použití jiných nástrojů (dánský archiv není zpřístupněn a jako databáze je použita DerbyDB)

Verze 3.10

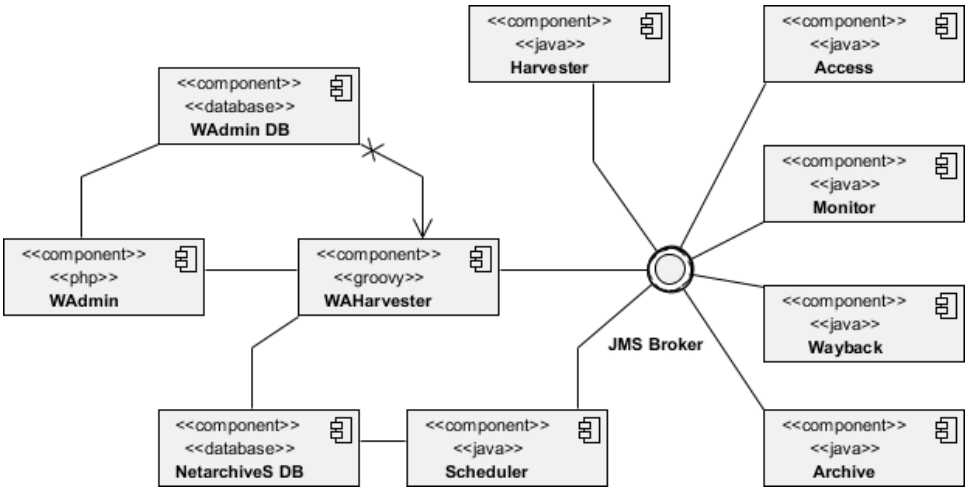
- Zapojení francouzské a rakouské národní knihovny do vývoje;
- přibyla lokalizace francouzštiny a italštiny;
- opravena chyba v použití databáze MySQL;
- deduplikace se stala volitelnou a lze ji kdykoliv vypnout;

1. Dokumentace systému NetarchiveSuite je k dispozici na adrese `http://netarchive.dk/suite/`.

- přibyly různé přehledy (např. všechna semínka sklizně, možnost filtrace atd.);
- podpora Waybacku.

Verze 3.12.1

- Nyní je možné ukládat pasti, které jsou globálně použity při všech sklizních;
- došlo k přepracování správy modulu BitArchive[2], informace se ukládají do databáze (implicitně je použita DerbyDB).



Obrázek 5.1: Možná integrace WA Harvesteru a Netarchive Suite

V diagramu 5.1 je znázorněn nástin možné integrace WA Harvesteru s NetarchiveSuite, kdy WAH komunikuje s JMS brokerem a sdílí databázi NetarchiveSuite. Tento přístup je sice reálný, ale pro současné požadavky až příliš komplexní a má i několik dalších nevýhod:

- Systém používá Heritrix 1, tato verze již dále není vyvíjena;
- systém je zbytečně komplexní v tom smyslu, že umožňuje individuální časování sklízení jednotlivých domén;

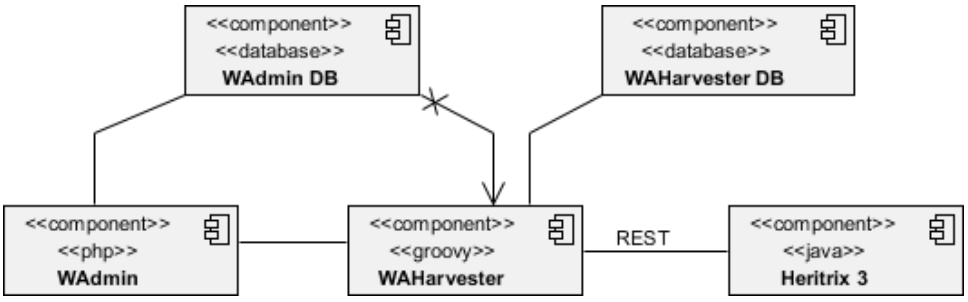
2. Tento modul zabezpečuje integritu dat a případně je možné ho zrcadlit na různé lokace.

- nelze provádět oddělenou indexaci QA a testovacích sklizní.

Z těchto důvodů nebude použita těsná integrace s NetarchiveSuite, ale při návrhu databáze bude brán zřetel na případné budoucí napojení systému NS. Proto jsem při tvorbě konceptuálního modelu korigoval část názvosloví podle existujícího datového modelu NS[3].

5.2 Heritrix 3

Heritrix 3 je novou verzí sklízecího robota, která dosud není v projektu nasazena v produkčním prostředí. Je ale jisté, že postupem času, a po migraci existujících profilů ze stávající verze, bude Heritrix 3 použit. Díky tomu, že tato verze dokáže sklizně provádět opakovaně v jednom adresáři (verze 1 vytvářela pro každou sklizeň nový adresář nehledě na příslušnost k určitému profilu) je integrace s Heritrixem 3 vhodným řešením. Zároveň použitá architektura *REST* umožňuje jednodušší a čistější vzdálený přístup. Robota lze na dálku ovládat pomocí REST příkazů GET a POST, kdy je zaslán příkaz v XML formě a ve stejném formátu je obdržena odezva od robota. Návrh architektury tohoto propojení je znázorněn v diagramu 5.2.

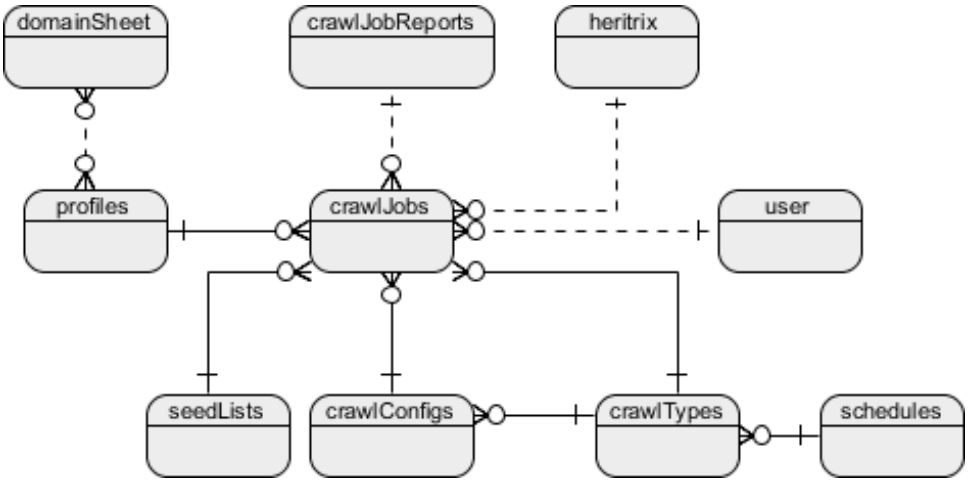


Obrázek 5.2: Možná integrace WA Harvesteru a Heritrixu 3

5.3 Konceptuální datový model

Aplikace bude obsahovat informace o objektech, které vznikají nebo mají souvislost se sklízením. Údaje o těchto objektech je nutné ukládat spolu s vzájemnými relacemi do databáze. Prvotní návrh konceptuálního modelu (tedy modelu znázorňující strukturu na nejvyšší úrovni) je uveden níže.

3. http://goo.gl/aONIT



Obrázek 5.3: Konceptuální datový model

5.3.1 Popis entit

V následující části jsem se soustředil převážně na nejdůležitější znaky entit a je pravděpodobné, že ve fázi implementační budou atributy doplněny.

crawlJob

Centrální entita, která obsahuje informace o provedených, probíhajících i plánovaných sklizních. Sklizeň musí být pojmenována a nachází se v určitém stavu, tyto stavy lze nalézt v dokumentaci Heritrixu 3 . U každé sklizně je uchován celý soubor nastavení (order.xml), priorita (značící důležitost sklizně a zařazení do fronty), název adresáře (kde je sklizeň vytvořena) a krátký popis.

Sklizeň je určitého typu (#crawlType), podle kterého je při založení nové sklizně určen konfigurační profil (#profile). Následně je po úpravách uživatelem vytvořen seznam semínek (#seedList) a konfigurace (#crawlConfig). Při spuštění samotné aktivity sklízení je zapsán uživatel (#user) a Heritrix (#heritrix), na kterém sklizeň poběží. Po skončení aktivity je Heritrixem vygenerován report dané sklizně (#crawlJobReport).

crawlConfig

V této entitě se nachází přesná konfigurace pro jednu sklizeň (#crawlJob). Tato konfigurace je vytvořena po založení sklizně a úpravě vybraného profilu (#profile). Konfigurace je možné znovu používat, dovolují nám tedy neopakovat stejné manuální úpravy a přitom nedochází ke změně základních profilů.

crawlType

Tyto entity reprezentují jednotlivé typy sklizní (#crawlJob), které jsou v projektu prováděny. Bude zde například typ "Výběrová sklizeň" nebo "Celoplošná sklizeň". Všechny typy mají přiřazen určitý časovač (#schedule), pomocí kterého lze v každém časovém okamžiku realizovat výběr nejbližších plánovaných sklizní (to bude provedeno výběrem typu a porovnáním posledních uskutečněných sklizní).

crawlJobReport

Každá z těchto entit bude obsahovat statistické údaje přesně pro jednu sklizeň (#crawlJob). Tyto údaje jsou dostupné po skončení sklízení a jsou získávány ze souboru crawl-report.txt, obsaženého přímo v adresáři sklizně. Obsahují počet a celkovou velikost dokumentů sklizených i čekajících ve frontě, dobu trvání, počet domén aj.

profile

Pro každý typ (#crawlType) sklizně existuje základní profil, ze kterého jsou odvozeny konkrétní konfigurace jednotlivých sklizní. Tento profil je upravován až na základě významnějších zjištění (přechod na nový server) a změn v rozhodnutí managementu (např. ignorování pravidel robots.txt).

domainSheet

V Heritrixu 3 je zaveden pojem domainSheet, což je sada pravidel, která jsou uplatněna jen pro určité domény. To poskytuje mocný nástroj pro selektivní úpravu sklizní. Pokud si vydavatel například nebude přát sklízet část stránek, či snížit rychlost, lze nadefinovat tyto pravidla do těchto specializovaných nastavení a opakovaně je použít. Systematická správa těchto úprav je velmi důležitým přínosem pro projekt, protože umožňuje vést historii v odlišených nastavení archivace jednotlivých domén. Tímto způso-

bem lze zároveň definovat i seznam globálních sklízecích pastí (kalendáře ve standardních blogovacích systémech apod.)[4].

seedList

Pro každou sklizeň (#crawlJob) existuje seznam semínek, která jsou určena k archivaci. Seznam bude uložen ve stejném formátu, jako je použit v Heritrixu. Uložen bude jako jednoduchý text, kdy na každém řádku je jedna doména, ta může být reprezentována jak *SURT* prefixem[5] tak v normální URL formě. Nutno podotknout, že v Heritrixu 3 je možné zapisovat semínka s prefixem plus (doména je do sklizně zahrnuta; jedná se o standardní chování, které je aplikováno, pokud prefix neuvedeme) nebo také s prefixem mínus, který značí vyloučení ze sklizně. Především prefix mínus v kombinaci se SURT prefixem domény je velice mocný nástroj pro vyloučení částí stránek.[6]

schedule

V této entitě je uvedena množina možných časování typů sklizní (#crawlType). Časy jsou ve formátu počtu dnů a obsahují popis časovače.

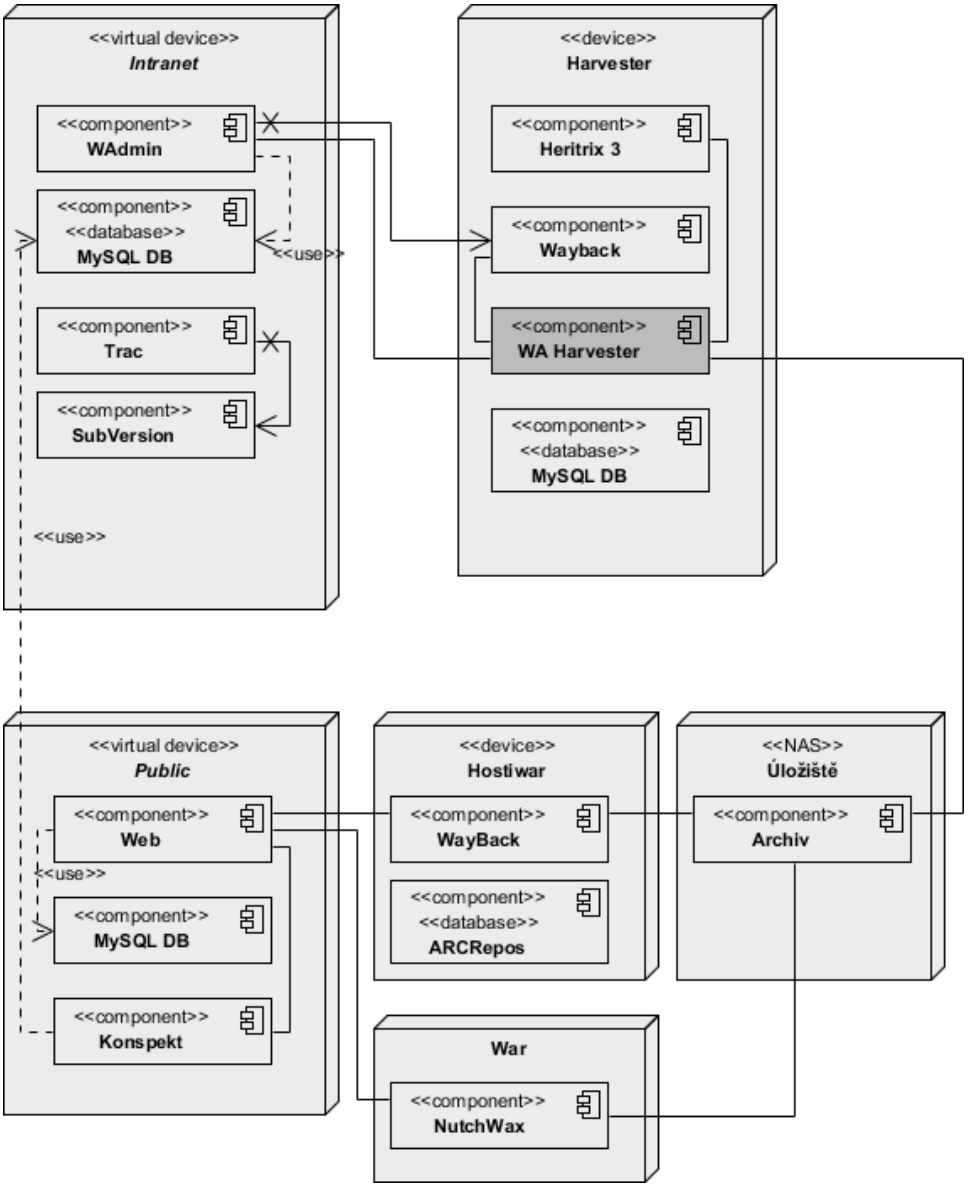
user

Pro každého uživatele obsluhujícího systém existuje v této tabulce záznam, který osobu jednoznačně identifikuje podle uživatelského jména a hesla. Je zde kontaktní email a celé jméno.

---

4. Zvážil jsem i variantu oddělené tabulky určené pouze pro pasti, ale zavrhl jsem ji jako zbytečnou redundanci dat, protože domainSheets jsou schopny tuto informaci reprezentovat.
5. SURT je zkratkou Sort-friendly URI Reordering Transform a jedná se o transformaci aplikovanou na URI, která obrátí pořadí jednotlivých částí adresy. Například URI <scheme://domain.tld/path?query> je transformována na SURT formu <scheme://(tld,domain,)/path?query>.
6. Například semínko "-http://cz,domena,eshop" zapříčiní vyloučení poddomény eshop.domena.cz

5.4 Návrh architektury



Obrázek 5.4: Deployment diagram - návrh architektury

5.5 Návrh uživatelského rozhraní

Jako každá jiná webová aplikace, i WA Harvester musí mít jednoduše použitelné uživatelské rozhraní. Aplikaci lze podle jednotlivých modulů rozdělit na logické sekce (například správa sklizní, statistky atd.) Tyto sekce jsou dále děleny pomocí činností, které k dané části přísluší (např. statistiky -> zobrazit roční statistky). V horní části obrazovky jsou umístěny základní informace – který uživatel je přihlášen, možnost odhlášení a na jakém serveru se aplikace nachází. Níže je menu, které obsahuje základní sekce a odkazy do systémů Trac a WA Admin, protože lze předpokládat časté prolínání informaci z jednotlivých aplikací. V levém menu bude logo, sloužící jako odkaz na hlavní přehled (dashboard), níže bude vyhledávací pole a seznam podsekcí či činností pro danou sekci systému. Ve vymezeném zbylém prostoru se bude nacházet samotný obsah podsekce. Pro jasnější a přehlednější definici jsem vytvořil diagram návrhu uživatelského rozhraní (Diagram 5.5).

Obrázek 5.5: Návrh uživatelského rozhraní

Kapitola 6

Programové řešení systému WA Harvester

6.1 Výběr programového prostředí

Systém WAH je realizován jako webová aplikace, která umožňuje správu a manipulaci s daty obsaženými v databázi a zároveň komunikaci se sklízecím robotem. Dále jsou uvedeny požadavky, které jsem zvažoval při výběru technologií a nástrojů použitých v systému WAH.

6.1.1 Základní požadavky na technologie

- Plná podpora programovacího jazyka Java – Heritrix, WayBack a velká část nástrojů v projektu je napsána v Javě;
- podpora REST – Heritrix je přístupný a ovladatelný pomocí REST architektury;
- podpora MySQL databázového stroje – MySQL je v rámci projektu WebArchiv stěžejní databáze;
- možnost tvorby testovacích případů;
- licence otevřeného zdrojového kódu;
- stabilita a bezpečnost;
- ideálně ustálený produkt, který má silnou komunitu.

6.1.2 Webový framework

Svět Javových webových frameworků je velmi dobře znám svojí roztříštěností a přemírou komplexity, proto je dnes velmi obtížné vybrat framework, ve kterém aplikaci budeme psát. Vybral jsem tedy několik významných frameworků a provedl jejich analýzu. Jsou jimi: *Spring*, *Struts 2*, *Wicket*, *Vaadin*, *Tapestry*, *Google Web Toolkit* (*GWT*) a *Grails*.

Při hodnocení jsem zvolil vylučovací metodu. Nejprve jsem vyloučil frameworky, které se zabývají primárně uživatelským rozhraním a AJAX integrací, protože tyto komponenty nebudou v novém systému stěžejní. Tím jsem vyřadil Vaadin a GWT. Struts 2 je poměrně komplexní bez jednoznačného přínosu a podobně je na tom Wicket, který je dynamičtější, avšak chybí mu propracovaná dokumentace. Tapestry umožňuje rychlejší vývoj, protože dokáže za běhu kontrolovat změny v kompilovaných třídách a následně je aktualizovat v aplikaci. Tento framework jsem nezvolil, protože v posledních verzích došlo k výrazným strukturálním změnám a budoucnost aplikace je v současné chvíli málo předvídatelná.

Spring je v prostředí Javy velmi významný pojem a po odkoupení firmou VMware lze očekávat ustálení vývoje. Zároveň je velmi nepravděpodobné, že tento projekt bude v blízké době ukončen. Je dostupná velice kvalitní dokumentace, návrh je robustní, avšak zároveň je Spring pro mé účely příliš složitý. Tuto komplexnost zjednodušuje framework Grails, který je postaven nad Springem, využívá tedy celou škálu již stabilních technologii, například *Hibernate 3* pro reprezentaci datové vrstvy a *SpringMVC* pro MVC rámec.

Po zhodnocení všech faktorů jsem se rozhodl pro Grails, protože umožňuje využívat všech výhod Javy a zároveň přispívá flexibilitou programovacího jazyka Groovy. Dalším faktorem je, že Heritrix 3 část své funkcionality staví nad rámcem Spring, který je s Grails úzce propojen. V tabulce 6.1 je možno nalézt přehled hodnocených frameworků.

	AJAX	MVC	Testy	ORM	Bezpečn.
Grails	ANO	ANO	Unit, Integration, Functional testy	GORM, Hibernate	Spring security
Spring	ANO	ANO	Mock a Unit testy	Hibernate	Spring security
Vaadin	GWT	NE	ANO	ANO	NE
GWT	ANO	NE	Selenium	NE	NE
Struts 2	ANO	ANO	Unit testy	ANO	NE
Wicket	ANO	Event-Driven	Mock a Unit testy	ANO s pluginem	ANO
Tapestry	ANO	ANO	Částečné	Hibernate	ANO

Tabulka 6.1: Porovnání vybraných frameworků

6.1.3 Grails

V této části popíši základní znaky vybraného frameworku,

- Napsán v *Groovy*, postaven nad frameworkem Spring;
- ctí zásadu "konvence před konfigurací";
- obsahuje objektově relační mapovaní (ORM) – použitím vrstvy Hibernate;
- umožňuje jednoduchý systém psaní pohledů (views) – stránky GSP (podobně jako JSP);
- architektura odpovídá MVC a je postavena na Spring MVC;
- ovládání frameworku je možné pomocí skriptu gant;
- v distribuci je přímo obsažen integrovaný Tomcat pro usnadnění vývoje;
- závislosti jsou automaticky kontrolovány pomocí Apache Ivy (podobně jako Maven)

- integrovaná podpora překladů pomocí i18n;
- systém má otevřený zdrojový kód a probíhá neustálý vývoj a zlepšování[1].

6.1.4 Groovy

Protože Grails jsou napsány v programovacím jazyku Groovy, krátce zde zmíním pár základních rysů. Jedná se o agilní dynamický programovací jazyk pro Java Virtual Machine (JVM), který je postaven na základech Javy. Je však inspirován také Pythonem, Ruby a Smalltalkem. Podporuje uzávěry, anonymní funkce, statické i dynamické typování. Důležitá je bezproblémová integrace s existujícími třídami a knihovnami Javy a stejně, jako aplikace napsané v Javě, se kompiluje přímo do Java bajtkódu, který je možné spustit kdekoliv, kde je JVM

6.2 Implementace

Při samotném programovém řešení jsem použil metodu *programování řízené testy* (*TDD*) [21]. Při této technice jsou nejdříve napsány testovací případy pro plánovanou funkcionalitu a teprve následně je implementována samotná funkcionalita. Testy jsou částí specifikace a lze z nich vyčíst očekávané chování programu.

Pro snadnou a udržitelnou správu kódu jsem použil systém *GIT*. Tento systém správy verzí je distribuovaný a poskytuje širší možnosti než v projektu WebArchiv používaný *Subversion*. Projekt jsem vytvářel pod licencí otevřeného kódu, konkrétně *Apache License 2*[2], a repositář je zveřejněn na serveru *GitHub*, kde jsou i případné informace o vývoji[3].

Rámec Grails používá architekturu MVC (Model–View–Controller), a proto v následující části stručně přiblížím jednotlivé komponenty. Aplikace je rozdělena na dva hlavní balíky, v nichž jsou logicky odděleny části, které zpracovávají požadavky pro WAH a ty, které spolupracují s aplikací WA Admin.

6.2.1 Modely

Modely (Models) reprezentují data, která jsou spravována aplikací. Řídí chování jednotlivých objektů obsažených v databázi a odpovídají na dotazy

1. Více informací lze nalézt na http://www.grails.org/Documentation
2. http://www.apache.org/licenses/LICENSE-2.0
3. http://github.com/nanux/WAHarvester

(často pocházející z pohledů), které se týkají stavu objektů. Dále reagují na pokyny změny stavu (převážně od řadičů) a zajišťují korektní provedení těchto změn, aby nedošlo k narušení integrity modelu. V Grails jsou doménové objekty specifikovány jako jednoduché třídy class, které jsou umístěny v adresáři `domains`. V jednotlivých modelech jsou definovány atributy, které mají přiřazeny datový typ a název. Další důležitou součástí jsou omezení obsahu jednotlivých atributů, tzv. constraints[4]. Například jméno nesmí být prázdné, musí být kratší než 30 znaků a je unikátní v rámci modelu. Tato omezení jsou při tvorbě datového modelu transformována do integritních omezení a vlastností datových sloupců.

V balíku `cz.webarchiv.wah` jsou definovány datové entity, které jsou specifikovány v konceptuálním datovém modelu (Diagram 5.3) a reprezentují data spravovaná systémem WAH. Každá doména obsahuje atributy, které vyplývají z popisu entit, a příslušná logické omezení. U některých atributů je změněno implicitní mapování, například atribut orderXML u sklizně je definován jako dlouhý text, protože základní typ VARCHAR by byl nedostačující. Nedílnou součástí domény je definice vazeb mezi modely, tyto relace jsou zapsány jako statické atributy belongsTo, hasOne a hasMany[5].

Pro zpřístupnění již existujícího doménového modelu, který je použit v systému WA Admin jsou v balíku `cz.webarchiv.wadmin` vytvořeny třídy, které obsahují stejné informace jako třídy z WAH, avšak jména atributů a omezení jsou přejata z databáze. Protože konvence Grails neodpovídají konvencím použitým ve frameworku *KohanaPHP*, bylo nutné nadefinovat správné mapování tabulky. Způsob napojení na databázi je popsán v [18].

6.2.2 Řadiče

Řadiče (Controllers) přijímají vstup od uživatele a na základě přijatých dat provedou úpravy v datovém modelu voláním metod příslušných objektů. Následně výsledek úprav zobrazí uživateli v pohledu. Řadiče tedy zajišťují transformaci uživatelských požadavků na odpovídající změny v modelu, mohou ale také fungovat jako správci pohledů, kdy není použit doménový model, ale pouze zobrazují pohledy podle požadavku uživatele. Každý řadič je pojmenován podle své funkce a sufixem Controller (CrawlJobController) a obsahuje metody, které definují možné akce. Grails interpretuje

4. Celkový přehled dostupných omezení lze nalézt na `http://www.grails.org/doc/latest/ref/Constraints/Usage.html`
5. `http://www.grails.org/doc/latest/guide/5.AssociationinGORM`

URL požadavku na správnou akci, například `/CrawlJob/list/10` – vyvolá metodu `CrawlJobController.list(10)`, která zobrazí 10 posledních sklizní.

Balík `cz.webarchiv.wah` obsahuje řadiče, které obsluhují požadavky systému WAH a jsou členěny na logické celky podle oblasti, kterou zpracovávají.

- Dashboard – zobrazuje přehled posledních akcí, upozornění a harmonogram.
- Monitor – zobrazuje informace o probíhajících sklizních a připojených sklízecích robotech.
- Statistics – umožňuje generování periodických i celkových statistik systému.
- CrawlJob – slouží ke správě, editaci a zakládání nových sklizní. Je zde také metoda importJobDirectory, která zajišťuje vkládání již existujících sklizní do systému.
- QualityAssurance – v tomto řadiči je možné přehledně zobrazit informace o provedených kontrolách kvality.
- CrawlProfile – slouží ke správě profilů jednotlivých typů sklizní.

Stejně jako v modelu, i u řadičů je vytvořen balík cz.webarchiv.wadmin, který umožňuje omezenou manipulaci s daty, uloženými v systému WA Admin.

6.2.3 Pohledy

Pohledy (Views) zobrazuqjí modely v uživatelském rozhraní a umožňují interakci s nimi. V systému Grails je možné psát pohledy v *JSP* a *GSP* (Groove Servers Pages), zvolil jsem flexibilnější formát GSP. Adresářová struktura je dána konvencí `/views/radic/akce`, tedy například pohled `/views/crawlJob/list/10` odpovídá příkladu řadiče uvedeného výše. Samotný obsah je složen z HTML kódu a GSP fragmentů, které zajišťují logiku při transformaci modelů a dalších informací z řadiče do zobrazitelné formy.

Protože Grails umožňuje použití šablonovacího systému Sitemesh[6], použil jsem obecnou šablonu, kterou lze nalézt v adresáři `/views/layouts/`

6. http://www.opensymphony.com/sitemesh/

a je standardně pojmenována `main.gsp`. Použití šablony snižuje nároky při tvorbě jednotlivých pohledů, protože informace, které se nejčastěji opakují, jsou uloženy na jednom místě a jsou automaticky vkládána do každého pohledu. Tato šablona vychází z návrhu *Adminizio Lite*[7], který je pro nekomerční účely volně dostupný.

6.3 Další služby

Pro zajištění některých služeb jsem vytvořil artefakty, které stojí mimo architekturu MVC a které popíši v následující části.

6.3.1 Import existujících sklizní

Důležitou součástí řešení je vytvoření modulu, který umožňuje importování již existujících sklizní. Tato služba se nachází v services a je pojmenována *ImportJobService*. Při zavolání metody `importDirectory(String path)` služba vyhledá vnořené sklizně v adresáři a následně zpracuje seznam semínek, nastavení sklizně (*orderXML*) a konečné výsledky sklizně. Takto zpracované sklizně jsou navráceny v seznamu jako instance třídy *CrawlJob*. Pro jednotlivé metody lze nalézt testovací případy, které jsou v současnosti testovány proti sklizním z Heritrixu verze 1 (tyto soubory lze nalézt v `/test/unit/resources`).

Příklad metody zpracovávající konečné výsledky sklizně.

```
protected CrawlJob parseOrderXML(File orderXMLFile) {
 if (orderXMLFile.exists()) {
  def orderXML = new XmlSlurper().parse(orderXMLFile)
  CrawlJob job = new CrawlJob()
  job.orderXml = orderXMLFile.getText()
  job.name = orderXML.meta.name.text()
  job.description = orderXML.meta.description.text()
  job.dateStarted = Date.parse(
                        'yyyyMMddHHmmss', orderXML.meta.date.text())
  job.totalBudget = orderXML.'**'.find
                        {it.@name == 'queue-total-budget'}
                        .text().asType(Integer)
  return job
 }
 else {
  throw new WaHarvesterException(
                'Can not open order.xml file: '
                 + orderXMLFile.getPath())}}
```

7. http://www.adminizio.com

6.3.2 Link extractor

Za účelem zjednodušení kontroly kvality jsem vytvořil[8] modul *Link Extractor*, který využívá knihovnu *HtmlUnit*[9]. Tato knihovna umožňuje přímo v kódu Java spouštět webový prohlížeč bez uživatelského rozhraní a v takto vytvořeném prostředí provádět typické akce, které jsou dostupné uživateli v prohlížeči (klikání na odkazy, vyplňování formulářů aj.) Protože knihovna podporuje v plné míře Javascript, AJAX, cookies a další komplexnější technologie, je mnohem úspěšnější v nalézání odkazů obsažených ve stránkách oproti metodě použité v Heritrixu (tj. analýze zdrojového kódu). V balíku cz.webarchiv.linkextractor se nalézá třída LinkExtractor, která obsahuje metodu *extractLinks*. Při vyvolání metody volající předá v parametru URL stránky určené k analýze. LinkExtractor stránku načte a zachytí každou značku `<A>` a `<IMG>`, ze kterých přečte atributy udávající URL odkazu. Dále pokud má tento objekt nadefinovánu událost onClick, provede virtuální kliknutí a zachytí URL, kam je prohlížeč přesměrován. Po analýze celé stránky vrátí kompletní seznam zachycených odkazů ve formě množiny (třída *Set*), kde každý záznam je URL v prosté textové formě. Tento způsob je úspěšný, navíc je zde velký potenciál pro další rozšíření schopností.

6.3.3 Napojení na WA Admin

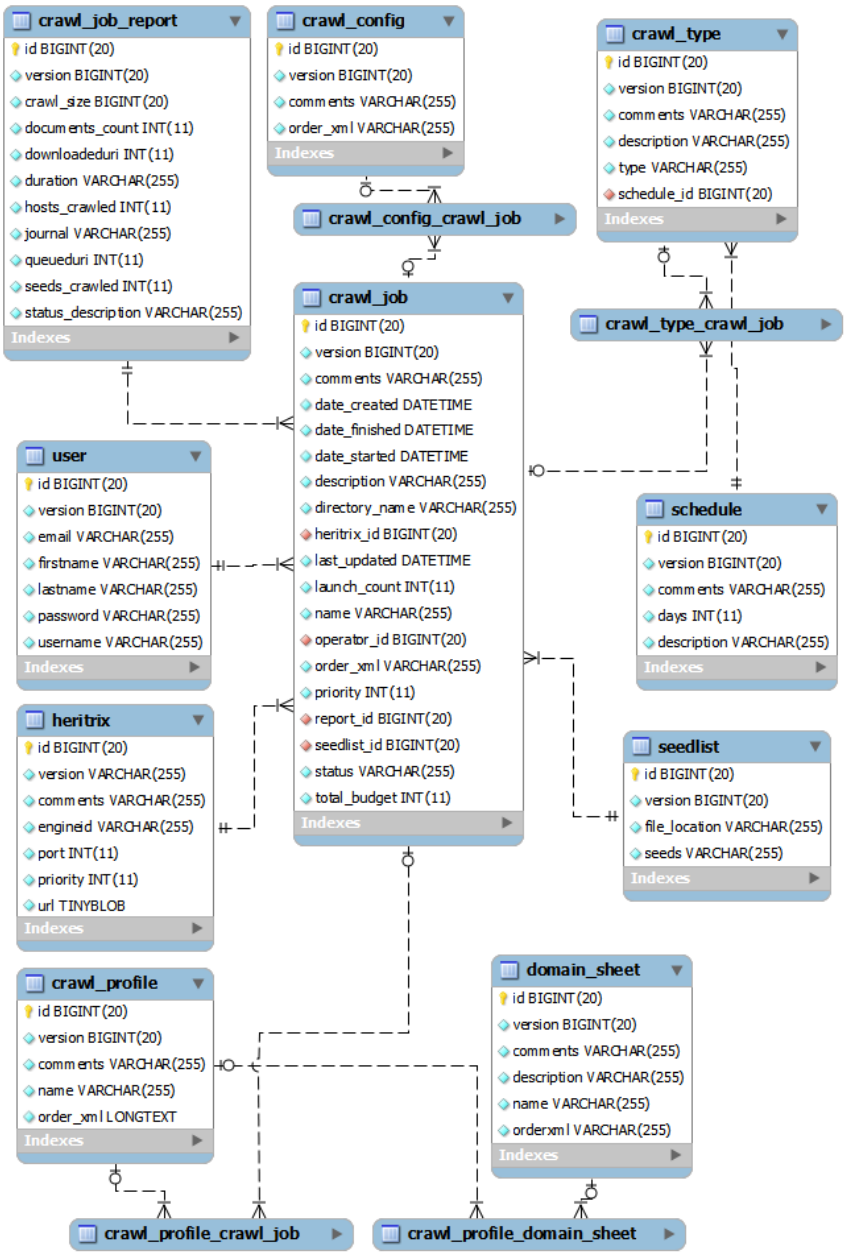
Protože systém Grails neumožňuje v základní verzi připojení k více než jednomu datovému zdroji, musel jsem pro připojení datového modelu WA Admin použít zásuvný modul *Datasources*[10]. V konfiguračním souboru `/conf/datasources.conf` je definováno připojení k druhé databázi a zároveň obsahuje seznam modelů, pro které má být toto připojení využito. V ostatních případech je použito standardně definované spojení.

```
datasources = {
    datasource(name: 'wadmin') {
        driverClassName('com.mysql.jdbc.Driver')
        dbCreate("update")
        url("jdbc:mysql://localhost/wadmin")
        username("user")
        password("heslo")
        domainClasses(
[cz.webarchiv.wadmin.Curator, cz.webarchiv.wadmin.Publisher,
cz.webarchiv.wadmin.Resource, cz.webarchiv.wadmin.Seed,
```

8. Modul vznikl spoluprací s Mgr. Vaškem Rosickým
9. http://htmlunit.sourceforge.net/
10. http://www.grails.org/plugin/datasources

```
cz.webarchiv.wadmin.SeedStatus, cz.webarchiv.wadmin.QaProblem,
cz.webarchiv.wadmin.QaCheck])
        dialect(org.hibernate.dialect.MySQL5InnoDBDialect)
        pooled(true)
        environments(['development'])    }}
```

6.4 Implementace datového modelu



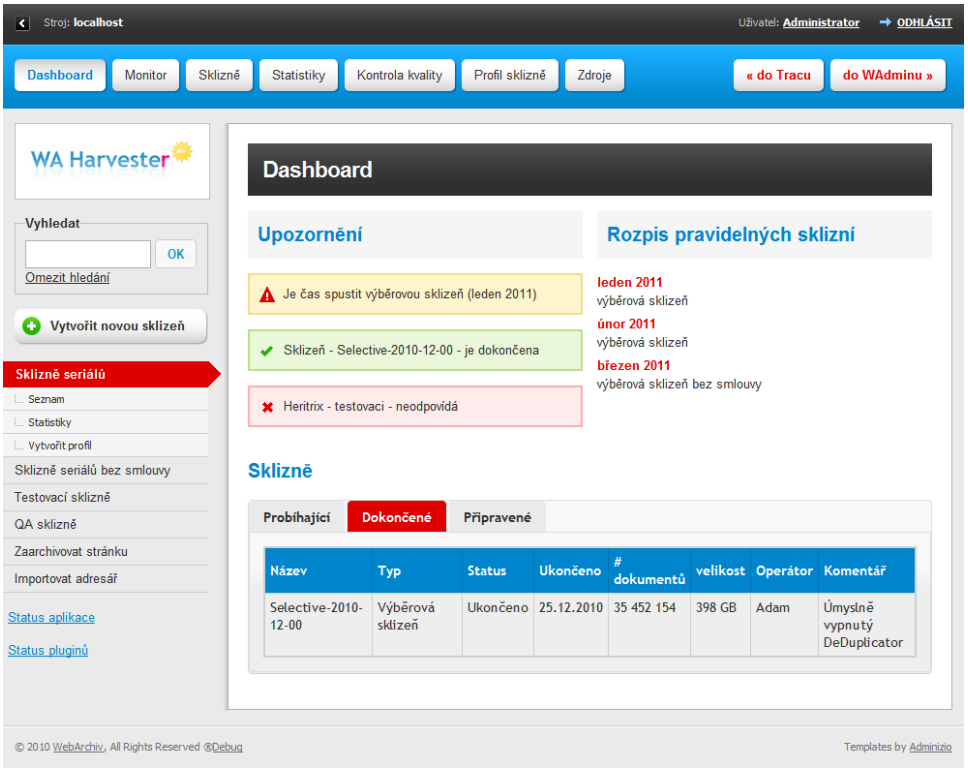
Obrázek 6.1: Implementace datového modelu

6.5 Nasazení systému WA Harvester

6.5.1 Popis nasazení

Aplikaci je možné ze zdrojových kódů vygenerovat příkazem grails war, po vykonání příkazu vznikne v adresáři target balík *WAR*, tedy standardní kontejnerový formát, který je použit pro Javové servery. Tento soubor (standardně pojmenovaný `WAHarvester-0.1.war`) stačí nahrát na server (testováno na Tomcat 5.5 a 6) do adresáře /webapps. Při standardním nastavení server zaregistruje nový balík a rozbalí ho do adresáře WAHarvester-1.0. Následně je aplikace přístupná na adrese `http://localhost/WAHarvester-1.0`.

6.5.2 Snímek obrazovky systému



Obrázek 6.2: Dashboard – hlavní přehled

Závěr

Vyvinutá aplikace umožňuje snížit časové náročnost sklízení dat a operátor sklizně se tím pádem může věnovat kreativnějším činnostem, jako je zajištění kvality archivace, případné další vývoj a optimalizace robota. Důležitým prvkem těchto automatizací je také systematická správa sekundárních souborů sklizně, které jsou nyní uloženy v jedné adresářové struktuře a zároveň jsou všechny potřebné informace vedeny v databázi. Existuje tedy možnost jednoduchého dohledání dat v případě potřeby, ať již kurátory, či managementem v NK. Tato databáze bude velice prospěšná i v případě přechodu na některý systém dlouhodobého uchování (LTP), který v budoucnu pravděpodobně proběhne. Nespornou výhodou, kterou aplikace přinese, bude přesné načasování sklizní, které je v současné době nutné hlídat manuálně. Již několikrát se v projektu sklizně plánované na určitý měsíc z důvodu chyby lidského faktoru opozdily, a nebyla tedy zachována plánovaná periodicita. Aby se toto riziko minimalizovalo, aplikace WA Harvester upozorní operátora na sklizně plánované v nejbližší době. Aplikace je nyní ve fázi důkladného testování před nasazením do produkčního prostředí.

Při zpracování práce jsem měl na paměti i skutečnost, že část procesů v projektu WebArchiv a aktuální technická specifika sklízení v ČR nejsou popsány. Práce tedy může sloužit i jako manuál pro zájemce, který se bude chtít o těchto detailech dozvědět více, případně se do projektu zapojit.

Existuje zde celá řada možností jak tento projekt dále rozvíjet. Poměrně komplexní problém automatizovaného ovládání sklízecího robota lze elegantně řešit pomocí složitějších REST dotazů, ale velkou obezřetnost je třeba vynaložit na chybové situace robota a zároveň předejít uživatelským zásahům, které by nechtěně vedly k havárii. Protože je ale Heritrix 3 šířen pod licencí otevřeného zdrojového kódu, lze v něm přímo vytvořit mechanismy, které budou kontrolovat správný chod a umožní těsnější integraci robota.

V projektu WebArchiv, ale i celosvětově, je patrný trend maximalizace kvality webového archivu. Pokud vezmeme v potaz rychlost vývoje webových technologií, je tento postup jistě správný. Velký praktický dopad by tedy měla formalizace procesu kontroly kvality sklizených stránek a s tím spojená implementace příslušných částí tohoto procesu do aplikace WA

Harvester, která by ulehčila práci kurátorům i operátorovi sklizně. Pro tyto změny je v aplikaci připraven datový model i speciální modul (Link Extractor).

Tvorba aplikace a této práce byla velice zajímavá a během teoretické části jsem prohloubil své znalosti o archivaci webu a v praktické realizaci jsem se naučil moderní metodiky, nástroje a zdokonalil své programátorské zkušenosti. Na aplikaci budu jistě dále pracovat a zdokonalovat ji.

Literatura

[1] A. Gulli and A. Signorini, "The indexable web is more than 11.5 billion pages." [online] 2005 [cit. 2010-01-07] Dostupný z WWW: <http://www.cs.uiowa.edu/~asignori/papers/the-in-dexable-web-is-more-than-11.5-billion-pages>.

[2] J. Alpert and N. Hajaj, "We knew the web was big...." [online] 2008 [cit. 2010-01-07] Dostupný z WWW: <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>.

[3] Kungliga biblioteket, "Svenska webbsidor - kulturarw3." [online] červen 2010 [cit. 2010-01-07] Dostupný z WWW: <http://www.kb.se/om/projekt/Svenska-webbsidor---Kulturarw3/>.

[4] J. Masanès, ed., *Web Archiving*. Berlin: Springer-Verlag, 2006.

[5] A. Brown, *Archiving websites : a practical guide for information management professionals*. London: Facet Publishing, 2006.

[6] R. Sanderson, L. Balakireva, H. Shankar, and H. V. de Sompel, "Transactional archives: A novel web preservation paradigm," tech. rep., Los Alamos National Laboratory Research Library, 2010. [online] [cit. 2010-01-07] Dostupný z WWW: <http://www.clir.org/dlf/forums/fall2010/16txn_archive.pdf>.

[7] G. Weikum, "Lawa: Longitudinal analytics of webarchive data." [online] 2010 [cit. 2010-01-07] Dostupný z WWW: <http://cordis.europa.eu/fp7/ict/fire/docs/fp7-factsheets/lawa_en.pdf>.

[8] Netarchive.dk, "Web archiving in Denmark August 2010 - a fact sheet." [online] 2010 [cit. 2010-01-07] Dostupný z WWW: <http://netarchive.dk/nyheder/Fact%20sheet%20Webarchiving%20in%20Denmark%202010.pdf>.

[9] PANDORA, Australia's Web Archive, "Pandora archive size and monthly growth." [online] 2010 [cit. 2010-01-07] Dostupný z WWW: <http://pandora.nla.gov.au/statistics.html>.

[10] National Library of Australia, "NLA guidelines for the development and application of a persistent identifier scheme for digital resources." [online] [cit. 2010-01-07] Dostupný z WWW: <http://www.nla.gov.au/initiatives/persistence/PIappendix1.html>.

[11] P. Žabička, "NEDLIB Harvester," *Ikaros*, 2000, roč. 4, č. 10. [online] [cit. 2010-01-07] Dostupný z WWW: <http://www.ikaros.cz/node/672>.

[12] A. Brokeš, "Projekt WebArchiv – archiv českého webu," *Zpravodaj ÚVT MU*, vol. roč. XVIII, no. 4, pp. s. 10–13, 2008. [online] [cit. 2010-01-07] Dostupný z WWW: <http://www.ics.muni.cz/zpravodaj/articles/578.html>.

[13] Creative Commons, "About the licenses." [online] [cit. 2010-01-07] Dostupný z WWW: <http://creativecommons.org/licenses/>.

[14] Internet Archive, "Users of Heritrix." [online] 2010 [cit. 2010-01-07] Dostupný z WWW: <https://webarchive.jira.com/wiki/display/Heritrix/Users+of+Heritrix>.

[15] A. Boyko, "Test bed taxonomy for crawler," tech. rep., IIPC. [online] 2004 [cit. 2010-01-07] Dostupný z WWW: http://netpreserve.org/publications/iipc-r-002.pdf.

[16] L. Matějka, "Zpřístupnění archivu českého webu," diplomová práce, Masarykova Univerzita, Fakulta Informatiky, 2006.

[17] R. Rivest, "The md5 message-digest algorithm," tech. rep., Network Working Group, 1992.

[18] A. Brokeš, "Systém pro správu procesu archivace webových informačních zdrojů," bakalářská práce, Masarykova Univerzita, Fakulta Informatiky, 2009.

[19] J. Arlow and I. Neustadt, *UML 2 a unifikovaný proces vývoje aplikací: objektově orientovaná analýza a návrh prakticky*. Brno: Computer Press, 2007.

[20] Projekt WebArchiv, "Kritéria výběru webových zdrojů." [online] [cit. 2010-01-07] Dostupný z WWW: <http://webarchiv.cz/kriteria/>.

[21] K. Beck, *Test Driven Development: By Example*. Boston: Addison-Wesley Longman, 2002.

Příloha A

Seznam webových archivů

Seznam institucí zabývajících se archivací Internetového obsahu:
Austrálie

- PANDORA Australia's Web Archive
- http://pandora.nla.gov.au/

Nový Zéland

- New Zealand Web Archive
- http://tinyurl.com/398qrh8

Evropa

- European Archive
- http://www.europarchive.org/

Česká republika

- WebArchiv
- http://webarchiv.cz

Velká Británie

- UK Web Archive
- http://www.webarchive.org.uk/ukwa/

- The UK Government Web Archive
- http://tinyurl.com/2u9r5j6

Island

- Icelandic Web Archive
- http://vefsafn.is/

Portugalsko

- Arquivo da Web Portuguesa
- http://www.arquivo.pt

Portugalsko

- UK Web Archive
- http://www.arquivo.pt

USA

- LC Web Archives Minerva
- http://tinyurl.com/c9gn3n

- Harvard's WAX
- http://wax.lib.harvard.edu/

- CDLIB
- http://webarchives.cdlib.org/

- Internet Archive
- http://archive.org/

Kanada

- Government of Canada Web Archive
- http://tinyurl.com/32wdvzv

Francie

- INA
- http://www.ina.fr/

Rozsáhlejší seznam lze nalézt na adrese `http://netpreserve.org/about/memberList.php`

Příloha B

Popis popis přiloženého CD

Na přiloženém CD je uložen výsledný systém a další praktické výsledky práce.

- WAHarvester.zip – komprimovaná aplikace WA Harvester, obsahuje zdrojové kódy a všechny potřebné artefakty pro spuštění i další vývoj aplikace.
- WAHarvester-0.1.war – WAR soubor, který lze nahrát na server a aplikaci nainstalovat.
- WAHarvester.sql – SQL skript zajišťující tvorbu struktury datového modelu (implicitně framework Grails tvorbu tabulek zajistí automaticky, skript je vytvořen pro úplnost).
- WAHarvester.vpp – zdrojový soubor pro aplikaci Visual Paradigm UML[1], který obsahuje vytvořené diagramy.
- WAHarvester.mwb – zdrojový soubor pro aplikaci MySQL Workbench[2] obsahující logickou strukturu implementovaného datového modelu.
- WAAdmin.sql – SQL skript zajišťující tvorbu datového modelu pro aplikaci WA Admin [3]
- /images – adresář obsahuje obrazové podklady pro aplikaci a diagramy.
- thesis.pdf – vysázený text práce v souboru PDF.
- thesis.tex – zdrojový kód pro sazbu práce ve formátu LaTeX.

---

1. `http://www.visual-paradigm.com/`
2. `http://wb.mysql.com/`
3. jedná se aktuální verzi z 9. 1. 2011