MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY

Systém pro správu procesu archivace webových informačních zdrojů

BAKALÁŘSKÁ PRÁCE

Adam Brokeš

Brno, 2009

Prohlášení

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

Vedoucí práce: RNDr. Miroslav Bartošek, CSc.

Poděkování

Shrnutí

Cílem práce bylo vytvořit náhradu za stávající systém, který je používán pro správu zdrojů, vydavatelů a smluv v projektu WebArchiv. Tato data jsou manuálně získávána a hrají důležitou roli ve výběrovém sběru elektronických dokumentů v rámci projektu. Bylo nutné vytvořit specifikaci pracovního procesu, navrhnout řešení a implementovat systém pomocí volně dostupných technologií. Specifikace je vytvořena v UML diagramech a programové řešení je implementováno v programovacím jazyce PHP spolu s využitím aplikačního rámce KohanaPHP. Datovou vrstvu zabezpečuje databáze MySQL a vývoj probíhal na serverovém stroji Apache 2.

Vysázeno v LaTeX

Klíčová slova

Archivace webu, WebArchiv, KohanaPHP, webový systém, sklízení, kulturní dědictví, elektronické dokumenty, autorský zákon, kurátor

Obsah

Úvod

Internet je médium, které původně vzniklo pouze pro výměnu informací mezi výzkumnými pracovníky, dosáhlo však v dnešní rozvinuté společnosti rozsahu masového média. Rychlost výměny informací, jeho přirozená svobodná podstata a množství obsažených dat, to jsou vlastnosti, které jsou mezi medií nepřekonatelné. Jeho rozvoj je spojen se zvyšující se kvantitou elektronicky zpracovávaných dokumentů. Tyto soubory nám přinášejí spoustu výhod. Můžeme s nimi jednoduše manipulovat, sdílet je a upravovat. Na druhou stranu to s sebou nese riziko ztráty či nechtěné změny. Životnost elektronických dokumentů je řádově mnohem kratší, než tomu je například u knih. Z tohoto důvodu vznikají projekty, které na sebe berou odpovědnost za zachování hodnotného obsahu, který bezesporu Internet obsahuje.

Jedním z nich je projekt WebArchiv. V této instituci probíhá sběr elektronických dokumentů dvojí cestou. V první je zachycen aktuální obsah celé české domény, v tomto případě je důležité především co nejširší pokrytí. Druhou cestou je pečlivý výběr kvalitních zdrojů informací, které následně podléhají velice podobným pravidlům jako vytištěné dokumenty. Každý zdroj je hodnocen týmem školených kurátorů, kteří rozhodují, zda-li ho zařadit do kulturního dědictví národa. Pokud tomu tak je, zodpovědný kurátor zašle oslovení vydavateli, ve kterém žádá o svolení zařadit zdroj do automatizovaně archivovaných dokumentů. V současné chvíli autorský zákon dovoluje data archivovat bez omezení, ale pro zpřístupnění je nutná smlouva. Pokud je vydavatelova odpověď kladná, zašle smlouvu a archivovaný obsah jeho webových stránek je zpřístupněn návštěvníkům.

Protože se nyní jedná řádově o stovky zdrojů a smluv, je nutné informace o nich systematicky udržovat. Celý projekt je zaměřen dlouhodobě a přírůstek nových zdrojů se stále zvyšuje. V takové situaci je systém spravující tyto informace velice důležitý. Protože však původní systém z mnoha důvodů již nevyhovuje, bylo nutné vyvinout systém nový, který bude vyhovovat nárokům kurátorů. Definování požadavků, vytvoření návrhu aplikace a jeho následná implementace je náplní této práce.

V první kapitole je rozvedena celá problematika archivace spolu s ná-

stroji, které se v tomto odvětví informačních technologií využívají. Druhá kapitola zkoumá systémy, které mají podobný účel a jsou zde uvedeny důvody, proč tyto nástroje pro potřeby Národní knihovny nevyhovují. Kapitola 3 obsahuje důležité části specifikace a analytické model, který je podkladem pro následující kapitolu, kde je navržena struktura aplikace a datového modelu. Závěrečná část popisuje implementační práce a architekturu aplikace.

Kapitola 1

Archivace webových informačních zdrojů

Webové médium patří mezi ta nejdynamičtěji se vyvíjející, a také ta nejkřehčí. Podle některých studií [19] je životnost elektronických dokumentů na internetu necelých sto dní. S přihlédnutím k tomu, že až 90 % těchto dokumentů existuje pouze v digitální podobě, není těžké si představit budoucnost, ve které naši potomci budou hledět na dnešní období jako na dobu „digitálního temna". Z těchto důvodů po celém světě vznikají instituce, které se zabývají archivací a zpřístupňováním (nejen) webových dokumentů. Protože tato činnost je analogická klasickým knihovnickým službám, vytvořila se specializovaná oddělení zabývající se danou problematikou především v rámci řady národních knihoven. Ze soukromých institucí jmenujme Internet Archive [1] - tento archiv shromažďuje data již od roku 1996 a velikost se pohybuje v řádech desítek petabytů. Všechny tyto instituce spojuje konsorcium IIPC [7] - *International Internet Preservation Consortium*, které koordinuje spolupráci mezi jednotlivými členy. Národní knihovna České republiky je členskou organizací od počátku roku 2007.

1.1 Projekt WebArchiv

Úlohou projektu *Webový Archiv* [16] je řešení problematiky archivace národního webu, tj. bohemikálních dokumentů zveřejněných v prostředí sítě Internet[1]. Jde o shromažďování webových zdrojů, jejich archivaci, ochranu a zajištění dlouhodobého přístupu. Provádí se jednak kompletní plošná archivace, tj. automatický sběr „celého" českého webu, souběžně však probíhá i výběrová archivace (nejzajímavějších webových zdrojů vybraných na základě selekčních kritérií [11]) a tematické archivace (zaměřené na určité aktuální téma, např. volby, povodně apod.). V současné době je stav řešení na úrovni funkčního provozu s testováním nových funkcí. K převedení do plně rutinních činností je zapotřebí jednak podstatné navýšení financování projektu, jednak změny stávající legislativy (zejména autorsko-právní) tak,

1. V současnosti je *národní web* vymezen doménou .cz

aby umožňovala plné zpřístupnění archivovaných zdrojů.

1.1.1 Historie a současnost

WebArchiv vznikl v rámci programového projektu výzkumu a vývoje *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet* pod záštitou Ministerstva kultury ČR. Projekt je řešen od roku 2000 v Národní knihovně České republiky a financován téměř výhradně z grantové podpory. Spoluřešitelem odpovědným za informační technologie je Moravská zemská knihovna v Brně (MZK), externím spolupracovníkem je Ústav výpočetní techniky Masarykovy univerzity v Brně (ÚVT). Na programovém řešení se podílí tým studentů Fakulty informatiky MU. 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 [20], robotem vyvíjeným Helsinskou národní knihovnou. 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 došlo 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.

1.1.2 Pracovní postup

V současné době je workflow rozdělen na technickou a logickou část. Pracovníci v Národní knihovně zajišťují výběr a hodnocení zdrojů, jejich katalogizaci a kontaktování vydavatelů. Dále vytvářejí popisná metadata (Dublin Core Metadata[4]), jsou důležitým spojovacím článkem mezi vydavateli a technickou podporou v Brně, také vytvářejí podklady pro prezentaci projektu, především obsah pro webové stránky. V Praze je také umístěn server zpřístupňující archív a webový portál projektu. Tuto část procesu hlouběji rozvedu v kapitole 3.

2. Sklízení je automatizované shromažďování relevantních elektronických dokumentů
3. Crawler je softwarový robot, jehož primárním účelem je sklízení elektronických dat

Brněnská část týmu se stará o technické zázemí projektu. Jsou zde umístěny dva servery. Probíhá zde sklízení dat, provoz interního systému, vývoj a testování. Zároveň je třeba udržovat hardware, jeho provoz a provádět údržbu a lokalizaci použitého software.

1.1.3 Provedené sklizně, popis archivu

Celoplošné sklizně – tato sklizeň probíhá na celé doméně **.cz**, v současnosti je seznam domén druhé úrovně získáván od registrátora *NIC.cz*. Úkolem sklizně je zachytit co nejširší rozsah bohemikálních dokumentů. Z důvodu dosavadních kapacitních omezení se tato sklizeň provádí jednou ročně.

- 2001 – První pokus o provedení celoplošné sklizně pomocí jednoho serveru s páskovým robotem, sklizeň nedokončena díky technickým problémům. Hloubka zanoření[4] byla nastavena na 25 odkazů.
- 2002 – Sklizeň byla přerušena z důvodu záplav a fyzického zatopení serveru umístěného v Národní knihovně.
- 2004 – Sklizeň proběhla úspěšně, zastavena byla po zaplnění diskového prostoru. Hloubka zanoření byla 50 odkazů.
- 2005 – První sklizeň provedena pomocí robota Heritrix. Šlo hlavně o zátěžový test nového software. Sklizeň byla zastavena po havárii robota, která byla způsobena nedostatky tehdejší verze.
- 2006 – Sklizeň pomocí Heritrixu byla pozastavena po zaplnění diskového prostoru. Byl nastaven limit 100 MB na soubor a 5 000 dokumentů na server.
- 2007 – Bylo sklizeno 81,3 mil. dokumentů o celkové velikosti 3,6 TB. Vstupem bylo 320 tisíc domén druhé úrovně a celý proces trval necelý měsíc.
- 2008 – Tato sklizeň proběhla již na novém serveru při použití poslední verze Heritrixu 1.14.2 a vstupem bylo 480 tisíc domén. Celkově robot sklidil 3,9 TB dat a 78 mil. dokumentů. Bohužel se vyskytl problém s vyčerpáním paměti ve virtuálním stroji, ve kterém Heritrix běží.

Výběrové sklizně – tyto sklizně probíhají periodicky několikrát ročně[5] na základě výběru určitého zdroje, který splňuje selekční kritéria. Výběr

4. Hloubka zanoření v datové struktuře stránky je vzdálenost od počáteční url domény.
5. V roce 2008 bylo provedeno 6 sklizní s rozestupy dvou měsíců.

probíhá v Národní knihovně a posléze je kontaktován vydavatel zdroje, který, pokud souhlasí, podepíše smlouvu o zařazení zdroje do archivu. Sklizený materiál, který je již v archivu umístěn nebo do něj bude zařazen v budoucnosti, je poté možné legálně plně zpřístupnit. Těchto smluv je v současné době přes 650. Právě tomuto procesu se budu nadále v textu věnovat a zachycení procesu správy smluv a zdrojů na nich uvedených je primární funkcí vyvíjeného systému.

Tematické sklizně – při tomto druhu sklizně je zacílena množina stránek týkajících se zvoleného tématu. Dosud proběhly sklizně: Dalimilova kronika, Povodně 2002, Vysočina, Volby 2006, Prezidentské volby 2008, Praha olympijská, Nová budova Národní knihovny a Nová budova Národní technické knihovny

V současné době je v archivu celkově uloženo 13 TB nekomprimovaných dat, což činí přibližně 300 milionů dokumentu. Celých 70 % je tvořeno HTML soubory, které se dají velice efektivně komprimovat.

1.2 Právní rámec

Dle autorského zákona[6] je v současné době možné archivovat dokumenty bez povolení autora (s přihlédnutím na případné přetížení serveru), avšak plně zpřístupňovat archivovaná data lze pouze se svolením autora. Toho lze dosáhnout dvojí cestou. V prvním případě vydavatel podepíše tištěnou smlouvu a tu zašle Národní knihovně. Pokud dokument splňuje všechny náležitosti je zanesen do WA Adminu, systému sloužící na správu smluv a zdrojů (viz 1.4). Vydavatel může také publikované dokumenty vydávat pod licencí Creative Commons [3] a tím pádem není smlouva nutná. Tento případ ale není v současné chvíli příliš rozšířen. V obou případech jsou následně modulem AutoContractMarker (viz 1.3.3) označeny všechny dokumenty (i zpětně) patřící k danému zdroji.

Autorský zákon také poskytuje možnost zobrazovat plný rozsah archivu, ale pouze na vyhrazených terminálech umístěných v Národní knihovně. V případě, že návštěvník přistupuje mimo tuto vymezenou síť, je mu nabídnut pouze přehled časových verzí dokumentu obsažených v archivu s upozorněním, že pro plné zobrazení musí navštívit některý z terminálů.

6. zákon 37/1995 §3, 46/2000(tiskový zákon) a pozdější novelizace 398/2006

1.3 Současné nástroje

1.3.1 Heritrix

Heritrix [5] je open-source sklízecí robot (crawler), který je vyvíjen společností Internet Archive. Je velice modulární, rozšiřitelný a nezávislý na platformě (je napsán v jazyce Java). Skládá se z frameworku (samotného jádra programu) a modulů (frontiers, processors, scopes, filters). Samotné nastavení Heritrixu je vytvoření konkrétního zapojení a zřetězení modulů. Tímto řetězcem poté projde každý URI (Uniform Resource Identifier) a je zpracován podle zapojených modulů. V současné době je k dispozici verze 1.14.2, která se zaměřila na zkvalitnění ochrany před pádem do pasti (dynamicky generované stránky na kterých se může robot zacyklit) a deduplikaci již sklizených dokumentů.

1.3.2 Wayback

Wayback [13] je open-source aplikace vyvíjená v jazyce Java společností Internet Archive pro zpřístupnění archivovaných dokumentů koncovým uživatelům, která nahradila původní Wayback Machine použitý přímo na stránkách archive.org. Dokumenty jsou indexovány a zpřístupňovány pomocí URL. Je implementována podpora pro hvězdičkovou notaci a tak je možné vyhledávat pomocí jednoduchých výrazů. Systém může pracovat ve třech módech: Archival URL – systém pomocí Javascriptu[7] změní url odkazy na stránce tak, že odkazují zpět do archivu. Proxy – systém se chová jako proxy server, je obtížné měnit časové verze. Timeline – u serveru vždy zobrazí časovou osu

V přípravě je fulltextové vyhledávání a lokalizace. V tuto chvíli je Wayback využit ve WebArchivu pro zpřístupnění celého archivu, avšak pro zobrazení obsahu, na který není smlouva, je třeba přistupovat na terminálech v Národní knihovně (viz 1.2). [17]

1.3.3 AutoContractMarker

Tento modul sekvenčně prochází databázi archivovaných dokumentů a příznakem označuje takové objekty, pro které je podepsána smlouva a mohou být zpřístupněny bez omezení [2].

---

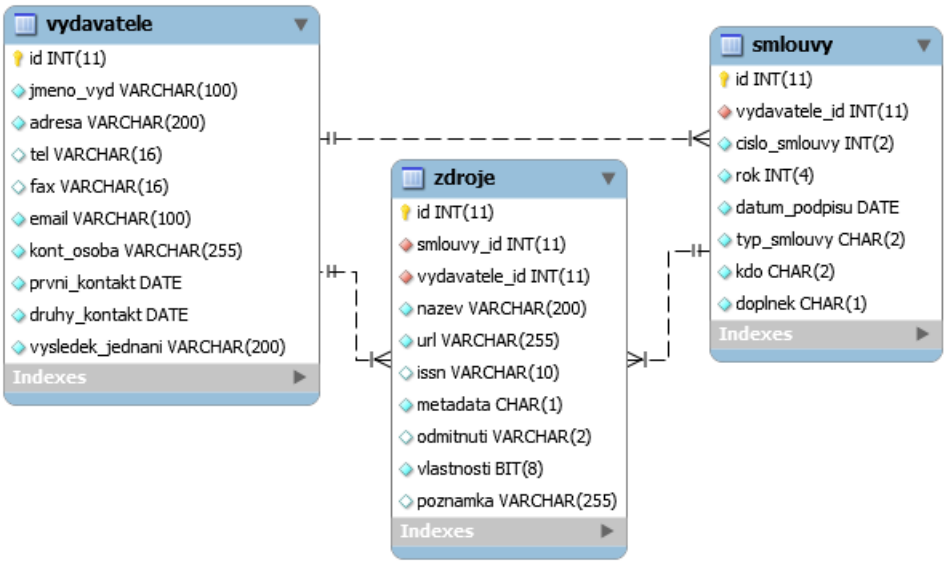
7. Javascript je skriptovací jazyk určený pro použití na straně klienta (webového prohlížeče)

1.4 WA Admin

Tento webový informační systém je určen především pro kurátory z Národní knihovny, kteří se zabývají získáváním a správou informací o vybraných zdrojích. Systém vznikl jako provizorní pomůcka pro kurátory v situaci, kdy nebylo nutné spravovat velké množství smluv a vydavatelů. Nepředpokládalo se, že by systém byl užíván po delší dobu[8] a proto při jeho tvorbě nebyly kladeny žádné nároky na rozšiřitelnost a možnost adaptace na případné změny. Systém je napsán ve skriptovacím jazyce PHP 4 a používá jednoduchý datový model implementovaný v MySQL (Obr. 1.1). Jeho funkcionalitu tvoří:

- správa zdrojů, smluv, vydavatelů
- zaznamenávání katalogizace a metadatového popisu
- poskytování statistik
- ze systému je také generován seznam spolupracujících vydavatelů, který je uveden na stránkách projektu[9]

1.4.1 Datový model



Obrázek 1.1: Původní datový model systému WA Admin

8. V současnosti je systém používán třetím rokem
9. http://webarchiv.cz/partneri/

Z aktuálních požadavků na systém vyplývá, že tento způsob uložení dat neobsahuje všechny nutné položky a spoustu důležitých informací je nutné uchovávat v poli *poznámka*, která je nevhodně umístěna pouze u zdroje.

1.4.2 Nedostatky

- zdroje nelze hodnotit
- datový model je příliš strohý (velkou část důležitých dat je nutné ukládat do datového pole poznámka)
- nezaručuje konzistenci uložených dat (používá MyISAM typ uložení dat, který je vhodný pro rychlé zpracování, ale nepoužívá cizí klíče)
- systém není flexibilní a neposkytuje ergonomické rozhraní

Navrhnout vyhovující náhradu za stávající systém je hlavním tématem této práce.

Kapitola 2

Možné existující řešení

Součástí práce je i prozkoumání existujících volně dostupných systémů, které mají podobný účel a vyhodnocení, zda-li jimi není možné nahradit původní systém.

2.1 NetarchiveSuite

Tento systém je vyvíjen pro potřeby Dánské národní knihovny [9]. Umožňuje automatizovanou sklizeň definovaných kolekcí URL. Tyto kolekce mohou být libovolně velké a tak je možné v tomto systému provádět celoplošné i výběrové sklizně. Zajímavým způsobem je řešena kontrola kvality sklizených dat, kdy kurátor prochází a porovnává archivovaná data s daty na živém webu a systém zaznamenává případné chybějící dokumenty, které sklidí při příští automatické sklizni. Implementace je provedena v jazyce JAVA a systém je dekomponován na množství nezávislých modulů, které komunikují pomoci JMS technologie. Sklízení pomocí tohoto software by bylo možné i technicky neerudovaným personálem.

2.1.1 Výhody

- sklízení se systémem je velice jednoduché
- poskytuje nástroje pro celoplošné, výběrové i tematické sklizně
- vytváří užší propojení mezi vybranými doménami a sklizeným materiálem, tím pádem je k dispozici okamžitá zpětná vazba
- části systému je možné distribuovat na více lokalit

2.1.2 Nevýhody

- nezahrnuje správu smluv, omezuje se pouze na domény

2.1.3 Závěr

Protože autorské právo v Dánsku se od českého výrazně liší v tom, že je zde dovoleno zpřístupňovat data pouze pro vědecké účely a nelze vytvářet výjimky ani prostřednictvím smluv, neobsahuje systém možnosti širšího popisu domén, ale omezuje se pouze na technické údaje. K datům nelze přikládat metadata v podobě smluv a implementuje tedy pouze sklízecí část pracovního procesu. Z tohoto důvodu je systém nevyhovující. Na druhou stranu je systém natolik dekomponován, že by bylo možné využít tohoto sklízecího modulu v nové verzi WA Adminu a zajišťovat tak automatizované sklízení.

2.2 Web Curator Tool

Web Curator Tool [14] je nástroj vyvinutý společným úsilím Národních knihoven Nového Zélandu a Británie v rámci projektu řízeného konsorciem IIPC. Systém podporuje správu smluvních zdrojů a následnou katalogizaci. Hluboko v logice programu je ustanoveno, že smlouvy je nutné uzavírat před samotnou sklizní, protože ta musí být také povolena vydavatelem. Důvodem je tamnější autorský zákon a z něho vyplývající workflow [15], které používají obě zainteresované instituce.

2.2.1 Výhody

- podporuje smlouvy v plném rozsahu
- práva jsou rozdělena mezi tzv. Agencies a to by umožňovalo separaci pracovního postupu mezi pracoviště (technické a kurátorské)
- kvalitní podpora historie provedených akcí
- přehledný systém sledování sklízení v reálném čase

2.2.2 Nevýhody

- nízká flexibilita systému
- přílišná robustnost a úzké propojení jednotlivých komponent
- neodpovídající workflow

2.2.3 Závěr

Naprosto zásadním problémem se ukázaly rozdíly v pracovním postupu. Zásada, že musí být smlouva ještě před samotným sklízením, je v přímém rozporu s analýzou požadavků na nový systém.

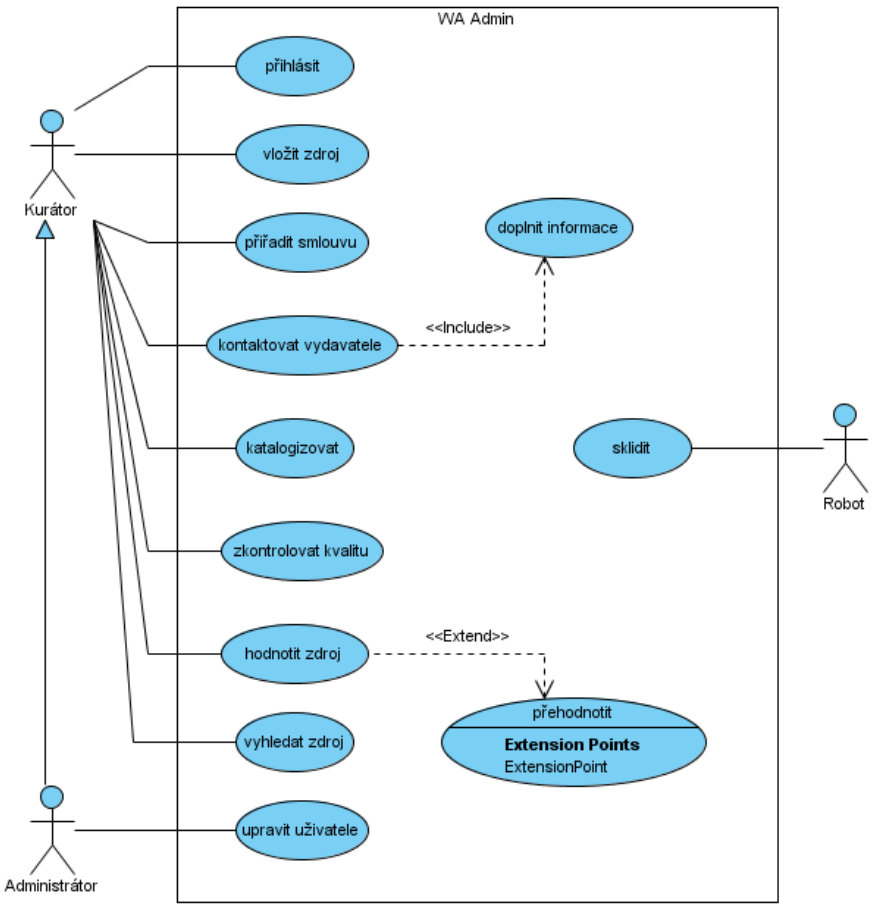
Kapitola 3

Specifikace systému WA Admin v2

V první fázi analýzy bylo nutné definovat a vymezit jednotlivé komponenty systému. K tomu jsem využil popis scénářů pracovního procesu. Na základě těchto informací a textového rozboru jsem sestavil analytický model, který jsem opět diskutoval se zadavatelem. Pro názornost byl doplněn stavovým diagramem nejkomplexnějšího objektu nalézajícího se v systému – zdroje. Dále jsem prozkoumal nároky na paměť a na rychlost, z důvodu správné volby datových struktur a úložišť ve fázi návrhu. Celá tato fáze je kompletně zdokumentována v příloze A.

Protože získání požadavků bylo třeba provádět s humanitně zaměřenými uživateli, zvolil jsem pro popis systému modelovací jazyk UML 2 [12]. Tato volba mi dovolila komunikovat s kurátory pomocí srozumitelných diagramů, které příliš nezatěžují technickými detaily.

3.1 Případy užití



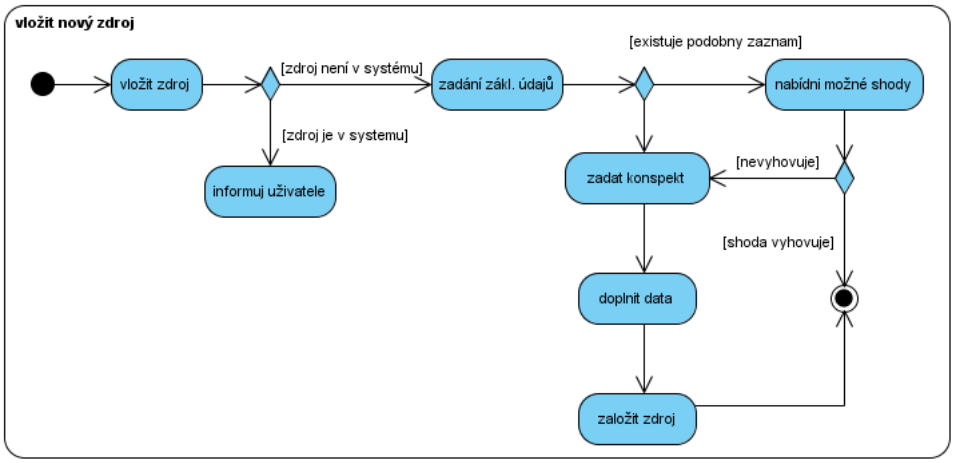
Obrázek 3.1: Případy užití systému

Ačkoliv samotné scénáře jsou ve většině případů vzájemně nezávislé, v následujícím textu budou řazeny v nejčastějším chronologickém sledu a budou uvedeny pouze netriviální a zásadní pro systém. V ostatních případech odkáži laskavého čtenáře na přiloženou dokumentaci (Příloha A).

3.1.1 Vložení zdroje

Před vložením nového zdroje systém ověří, že daný zdroj či vydavatel ještě není vložen. Kurátor zadá název zdroje, jméno vydavatele a URL. Systém

provede prohledání databáze na přítomnost některého z těchto údajů (logické OR). Pokud najde alespoň jeden z nich, upozorní kurátora a vypíše seznam potenciálně odpovídajících zdrojů nebo vydavatelů. Např. „Byly nalezeny tyto existující záznamy, které by mohly být shodné“. Pokud není nalezena potenciální shoda nebo nalezená shoda byla falešná, kurátor doplní další údaje a založí nový zdroj. Při prvním vkládání před hodnocením není třeba vkládat všechny údaje. Stačí vložit minimální základ – kdo zdroj vložil[1], název zdroje, jméno vydavatele a URL, kategorii Konspektu[2], typ návrhu, ISSN a datum vložení. Vydavatelé již vložení v databázi se nevkládají znovu, pouze se jim přiřadí nový zdroj.



Obrázek 3.2: Vložení nového zdroje - diagram aktivity

3.1.2 Hodnocení

Kurátor otevře seznam nově vložených zdrojů k hodnocení. Zobrazí se seznam názvů, URL, tlačítka pro jednotlivá hodnocení a textové pole pro případný komentář. Zároveň lze zobrazit kategorii Konspektu a odpovědného kurátora, pro případ, že se při hodnocení zjistí, že bylo původně přiřazeno nesprávně a je třeba provést změnu. K dispozici bude možnost „Zobrazit podobné“ a po jejím otevření se zobrazí seznam URL všech zdrojů z dané kategorie Konspektu.

1. Informace se doplní automaticky na základě přihlášeného uživatele.
2. Systém na základě Konspektu automaticky doplní odpovědného kurátora a jeho uživatelský kód

Po hodnocení všemi kurátory se přiřadí výsledek hodnocení – ANO, NE, MOŽNÁ (k přehodnocení za půl roku), TECHNICKÉ NE (zdroj je schválen, ale není technologická možnost ho v současnosti sklidit; pokud se tento stav v budoucnu změní, je osloven). Systém po ohodnocení všemi kurátory vypočítá konečné hodnocení, ale zároveň ponechá možnost hodnocení změnit. Finální hodnocení schvaluje správce záznamu.

Možné výsledky

- Ano – interval $\langle 1; 2\rangle$
  - zdroj je schválen a postupuje k oslovení
- Možná (přehodnotit) – interval $\langle 0,5; 1)$
  - zdroj je určen k přehodnocení
  - po šesti měsících se zobrazí na nástěnce všech uživatelů s příznakem „k přehodnocení“
- Ne – interval $\langle -2; 0,5)$
  - zdroj je zamítnut
- Technické ne
  - je zaznamenáno, pokud alespoň jeden kurátor ohodnotil zdroj jako TECHNICKÉ NE
  - je použito, pokud jsou předpokládány technické obtíže při sklízení, které znemožňují sklidit zdroj v požadované kvalitě
  - tyto zdroje je možné periodicky ověřovat

3.1.3 Přiřadit smlouvu

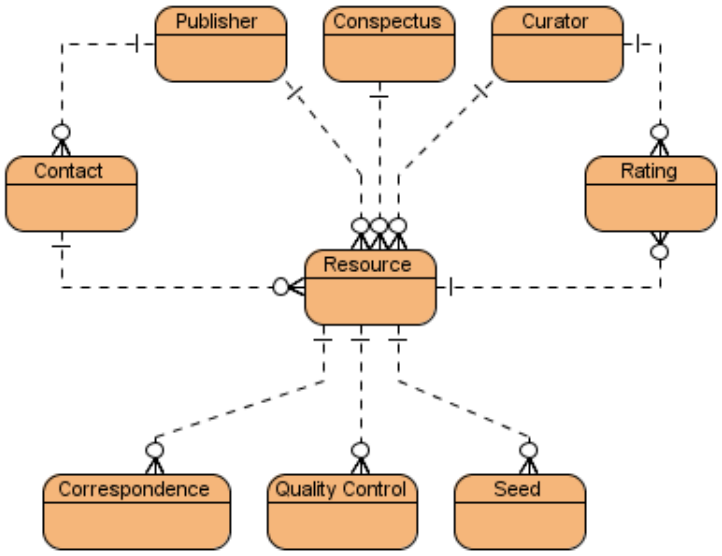
Pokud vydavatel zašle smlouvu o umístění zdroje do projektu WebArchiv, je ke zdroji přiřazena nová smlouva. Každá smlouva obdrží číslo generované systémem, skládá se z roku a pořadového čísla. Na jedné smlouvě může být uvedeno více zdrojů. V takovém případě je třeba přiřadit všem zdrojům z jedné smlouvy stejné číslo smlouvy.

Pokud vydavatelé uvedou do smlouvy více zdrojů, ale jedná se pouze o alias[3], je nutné tuto informaci evidovat pro zajištění indexace AutoContractMarkeru a umožnit přístup přes všechny tyto URL. Tyto zdroje se nezobrazují v seznamu spolupracujících vydavatelů na webu, který je gene-

3. Obsah webů je stejný a často všechna URL přesměrovávají na jednu z nich.

rován z WA Admin. Formulář pro smlouvu vychází z původního WA Admin, ale není použit typ smlouvy – je pouze jeden a smlouvy jsou ve tvaru – číslo smlouvy/rok – např.: 7/2008. V případě, že zdroj je vydán pod licencí Creative Commons, je vytvořena smlouva, u které je tato informace evidována.

3.2 Analytický model

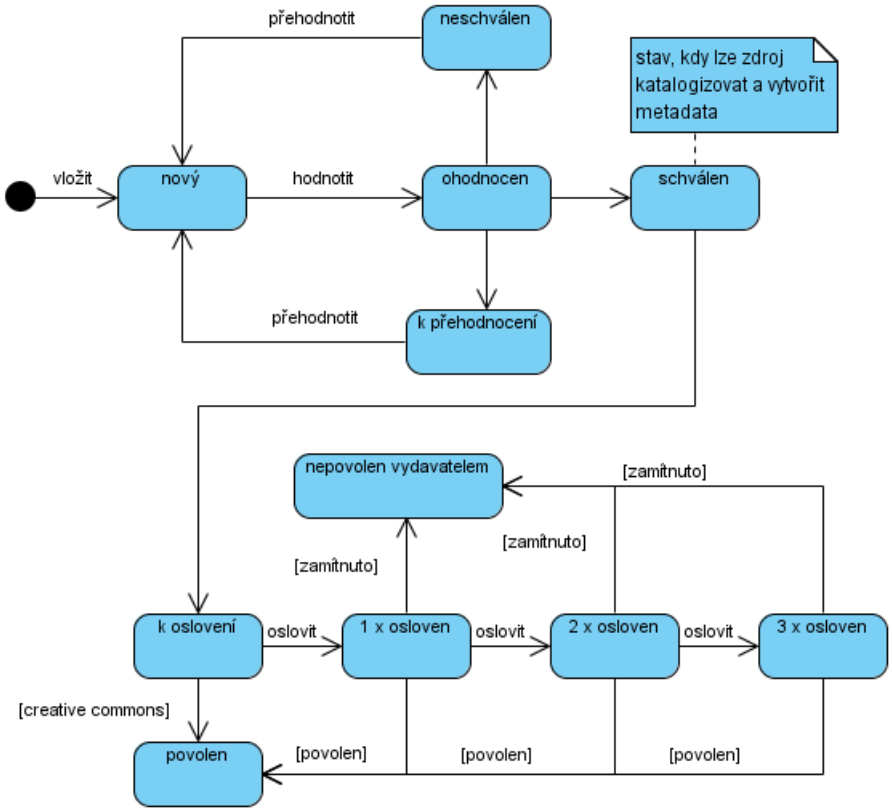


Obrázek 3.3: Analytický model

Na základě analýzy chybějících datových polí v původní verzi WA Adminu a případů užití jsem sestavil analytický model systému (Obrázek 3.3), obsahující pouze základní informace a vazby mezi entitami. Z něho je patrný stěžejní objekt systému a to jest zdroj. Ten je třeba chápat jinak, než jednu doménu s připojenými informacemi o vydavateli, jak tomu bylo v předchozí verzi systému. Zde se jedná o zapouzdřený objekt s abstraktním významem. Lze jím vystihnout v podstatě libovolně rozměrnou jednotku sklízených dat. Může se jednat o celý rozcestník webového portálu procházejícího přes více národních domén, stejně tak jako o jedinou stránku. Pro přesnější definici objektu jsem sestavil stavový diagram (Obrázek 3.4). Došlo k oddělení kontaktů od vydavatelů, vycházející z empirických zkušeností kurátoru, že vydavatel se mění výjimečně, zatímco kontakt poměrně

často a běžným případem je situace, kdy vydavatel má více zdrojů a pro komunikaci je vyhrazena vždy jiná osoba. Zcela novým prvkem, je systém hodnocení, které v modelu vytváří m:n relaci mezi kurátory a hodnocenými zdroji.

3.2.1 Zdroj - stavový diagram



Obrázek 3.4: Zdroj - stavový diagram

3.3 Nefunkční požadavky

Protože výběr zdrojů je pečlivá a manuálně náročná práce, není možné důležité části procesu automatizovat a tak rychlost bude určována především lidským faktorem. Důraz je tedy kladen více na zachování konzistence a integrity dat než rychlosti vkládání a výběru. Tyto časy se při správné dekompozici modelu budou pohybovat v přijatelných hodnotách.

Kapitola 4

Návrh WA Admin 2

V této fázi je do podrobností rozpracován analytický model. Jsou doplněny atributy jednotlivých entit spolu s datovým typem, případné výčtové typy atributů jsou dekomponovány na další tabulky. Druhá část obsahuje náčrt uživatelského rozhraní a jeho popis.

4.1 Datový model

Jak již byl zmíněno, nejrozsáhlejším objektem je zdroj, avšak z důvodu zachování flexibility je většina jeho položek nepovinná. Pohled na data v systému může být dvojí, data mohou být externí a interní, podle toho zdali jsou generována systémem. Celkový datový model je obsažen v přílohách (příloha B).

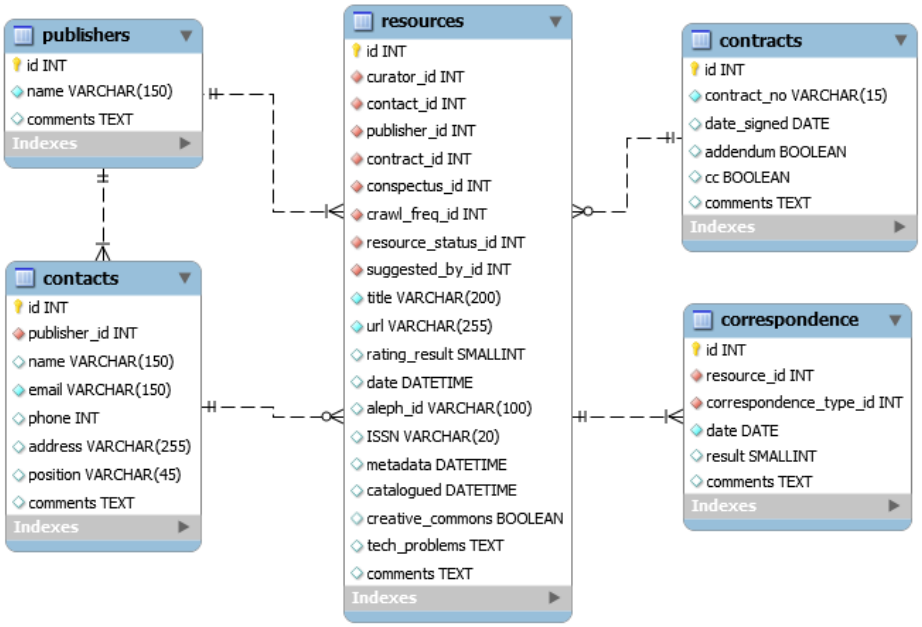
4.1.1 Externí informace

Zde jsou obsaženy všechna data, která mají původ vně systému. Jsou to tedy došlé smlouvy, zaslaná oslovení a informace o vydavatelích a jejich kontaktech. Tyto položky vyplňuje kurátor a je třeba pečlivě kontrolovat jejich syntaktickou správnost (např. email).

Každý zdroj je publikován vydavatelem se kterým kurátor komunikuje prostřednictvím kontaktu. Tomuto kontaktu zasílá různé typy oslovení, ve kterých nabízí vydavateli možnost zařazení zdroje do projektu WebArchiv. Odpověď vydavatele je zaznamenána jako výsledek oslovení. V případě úspěšného oslovení je následné obdržení smlouvy zaneseno do systému. Speciálním případem je zdroj pod licencí creative commons, pak je vytvořena nová smlouva s příznakem *cc* a zdroj je také označen touto licencí.

4.1.2 Interní informace

Zásadním požadavkem na systém bylo ukládání více URL adres určených pro sklízení (tzv. semínka) jednoho zdroje, tato možnost je zachycena zvlá-



Obrázek 4.1: Datový model - externí entity

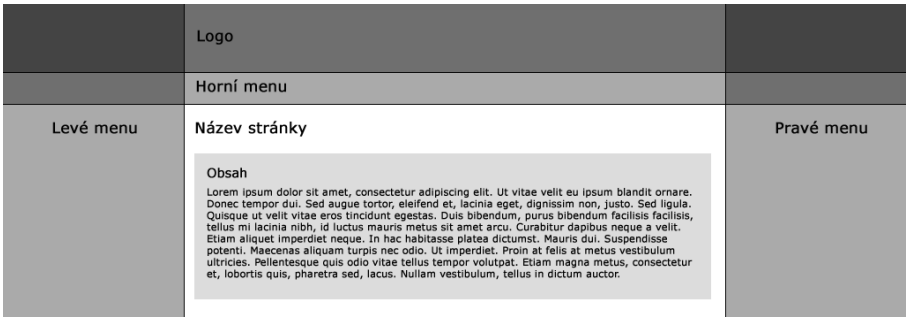
štní tabulkou semínek. Ta také dovoluje situace, kdy zdroj již není na adrese publikován, případně že se jedná o přesměrování. Všechny tyto vlastnosti jsou úzce propojeny se sklízením zdrojů pomocí Heritrixu. Každý kurátor je identifikován uživatelským jménem a do systému se přihlašuje pomocí hesla, které je hašované[1]. Systém může uživateli zasílat upozornění a informace na zadanou emailovou adresu a kolegové mají možnost zjistit i jiné kontakty (icq, skype). Hodnocení probíhá iterativně a každá iterace je uzavřena ohodnocením všemi uživateli, případně odsouhlasením správcem zdroje. Číslo iterace je uloženo v atributu *round*. Po prvním sklizení zdroje provede kurátor kontrolu kvality. Tato akce je nutná pro dodržení co nejvyšší kvality archivu a zpětné vazby mezi správcem sklizní a kurátory. Poté, co kontrolor prověří kvalitu sklizených dat, zanese odpovídající hodnocení do systému a v případě nesrovnalostí připojí komentář, který je vstupní informací pro změny v nastavení způsobu sklízení.

1. Heslo je zpracováno hašovací funkcí, např.: MD5.

Obrázek 4.2: Datový model - interní entity

4.2 Návrh rozhraní

Při základním náčrtu rozhraní bylo s uživateli ustanoveno, že ergonomicky nejvhodnější bude třísloupcové rozvržení s obsahem ve střední části.



Obrázek 4.3: Návrh rozhrani

Hlavní menu – kopíruje pracovní postup, musí tedy obsahovat položky:

- Hlavní stránka
- Vkládání zdrojů
- Hodnocení
- Oslovování
- V jednání
- Katalogizace

Pravé menu – z tohoto menu je možné se dostat na editor jednotlivých tabulek v databázi a zároveň slouží jako rozcestník pro interní aplikace, kterými jsou: wiki, blog a trac.

Levý sloupec – zde se nachází informace o přihlášeném uživateli a případná interaktivní nápověda k vybrané položce.

Kapitola 5

Implementace WA Admin 2

Volba implementačních technologií byla do určité míry determinována typem aplikace a provedenou analýzou a návrhem. Jako hlavní programovací jazyk jsem zvolil PHP [10], tento široce rozšířený multiplatformní jazyk byl stvořen pro webové aplikace a ve verzi 5 již plně podporuje standardy objektového programování. Je to jazyk skriptovací a tak je možné úpravy provádět za běhu aplikace bez nutnosti překompilování. Na druhou stranu jsou skriptovací jazyky v zásadě pomalejší, než jazyky kompilované. Tuto nevýhodu lze však minimalizovat použití vhodného aplikačního rámce a vhodným použitím eAcceleratoru. [18]

Jako databázový server jsem zachoval MySQL, protože s ním mají uživatelé již zkušenosti a je nejpoužívanější v celém projektu WebArchiv. Změnu jsem však provedl u způsobu uložení dat. Protože je kladen velký důraz na konzistenci a integritu, zvolil jsem InnoDB jako implicitní typ tabulek. Tento druh zajišťuje požadované vlastnosti již na datové vrstvě a předchází tedy omylům, které by mohly mít původ na aplikační úrovni. [6]

5.1 Model-View-Controller architektura

Základní myšlenkou této architektury[1] je oddělení aplikační logiky a prezentační vrstvy. Jak název napovídá, skládá se ze tří modulů: modelu, řadiči a pohledu . V modelu je jádro celé aplikace, obsahuje stav a data systému. Pokud dojde ke změně stavu modelu, nastane aktualizace všech závislých pohledů. Řadiče slouží klientovi k ovládání aplikace, jde o rozhraní k prezentační vrstvě, kterou tvoří pohledy. Každý pohled transformuje a zobrazuje data z modelu.

1. MVC se často se zaměňuje s návrhovým vzorem.

5.2 KohanaPHP

Základním kostrou celé aplikace je framework KohanaPHP [8]. Tento rámec vytváří plně objektové prostředí a jeho jádro tvoří MVC architektura. Systém vznikl přepracováním velice podobné aplikace – CodeIgniteru. Aby systém využil plného potenciálu páté verze PHP, byla přepracována jeho architektura do čistě objektové podoby. Systém má velice kvalitní knihovnu ORM[2]. Při použití tohoto nástroje se základní objekty modelu odvozují z tabulek databáze a my k nim můžeme přidat aplikační logiku a další atributy. Nejsme tím pádem nucení definovat duplicitně struktury a vazby, které jsou již obsaženy v datové vrstvě.

Velice zajímavou funkcí je kaskádovité překrývání objektů. V praxi to znamená, že KohanaPHP poskytuje širokou paletu knihoven a základních tříd. Pokud je ale některá struktura pro naše účely nevyhovující, můžeme ji jednoduše překrýt. Jedná se tedy spíše o předefinování chování celé vestavěné třídy, než o dědičnost. Ještě bych zmínil pomocný nástroj pro tvorbu formulářů FORGE (FORm Generator), který poměrně pohodlně vytváří webové formuláře a zajišťuje i následnou validaci. Zpracování dat už musí zajistit aktivní řadič.

5.3 Modely

Každý model musí být uložen v adresáři /application/models, jeho název je zakončen suffixem *_Model* a je dceřinou třídou Modelu (nemusí být bezprostřední rodič).

Table_Model

Tato třída je základem pro všechny modely odvozené od datového modelu. Sama je podtřídou ORM, takže obsahuje všechny metody a atributy pro přístup k datům, zjišťování přidružených objektů, vkládání, mazání řádků tabulky aj. Zavedl jsem ji z důvodu zvýšení flexibility, implementuje navíc chráněné atributy headers a default_column. První zmíněný je seznam položek, které se zobrazují v tabulkovém přehledu a druhý atribut je textovou reprezentací objektu. S druhým atributem souvisí i překrytá metoda toString(), která tento řetězec vypisuje.

2. ORM neboli objektově-relační mapování.

Název_Tabulky_Model

Každá tabulka obsažená v databázi má přidružený model odvozený od Table_Model. Zajišťuje přístup k jednotlivým atributům, zároveň je při volání speciální metody __set() zajištěna správnost dat.

5.4 Řadiče

Řadiče jsou reprezentovány webovou stránkou, která se skládá z pohledů a jsou primárním komunikačním kanálem mezi uživatelem a systémem. V našem případě je každý řadič obrazem logického kroku v pracovním procesu a většina odpovídá případu užití, které jsou popsány v dokumentaci. Nadtřídou těchto objektů je *Template_Controller*, který umožňuje použití implicitní šablony. Po přihlášení je vytvořen řadič *Home_Controller*, který získává data od modelu *Dashboard*. V něm má uživatel přehled o objektech, které čekají na manuální zpracování, ať už se jedná o hodnocení, katalogizaci, oslovení aj.

5.5 Pohledy

Slouží pouze k prezentaci uložených dat, případně předávají řadičům povely zadané uživatelem. Stránky jsem rozdělil na několik samostatných částí a pro každou jsem sestavil vlastní pohled, vznikla tak levá navigační lišta, kde je rozcestník k tabulkám, horní menu korespondující s kurátorovým postupem a pravá informační lišta. Výhodou je možnost zanořovat jednotlivé pohledy do sebe a tak mohla být vytvořena implicitní šablona složená z obsahově statických částí.

Pohledy jsou implementovány v jazyce XHTML s použitím kaskádových stylů. Barevné schéma jsem volil modro-oranžové pro maximalizaci ergonomie. Ukázku rozhraní je možné vidět v příloze C.

5.6 Pomocné třídy a rozhraní

5.6.1 Internacionalizace

Systém používá standardizovanou knihovnu i18n a umožňuje velkou část systémových výstupů jednoduše upravit a případně přeložit do cizích jazyků. Další výhoda spočívá v předání editačního nástroje uživatelům, kteří si jsou schopni modifikovat rozhraní a informativní zprávy.

5.7 Migrace existujících dat

Ačkoliv datový model původního systému byl malého rozsahu, obsahoval velké množství ručně vložených dat. Jednalo se o 2520 zdrojů, 2140 vydavatelů a 658 smluv. Tyto data nepodléhala žádné systematické kontrole a tak je možné v databázi najít množství defektů. Jako příklad mohu uvést více url adres v jednom záznamu, často chybějící informace o správci apod. Vytvořit migrační strategii a zajistit automatizovaný převod dat byl netriviální úkol, který je součástí této práce. Kvůli tomuto účelu jsem vytvořil dva nástroje. První z nich je SQL skript, který automaticky převádí data do nové databáze podle nadefinovaných závislostí. V druhém nástroji jsem zvolil opačný, tedy manuální přístup a vytvořil jsem komplexní dialog, zobrazující o zdroji pokud možno co nejvíce informací, které může kurátor editovat.

5.7.1 SQL skript

Činnost skriptu bych rozdělil na tři úrovně. V první jde pouze připravení existujících dat pro konverzi. Je nutné změnit datový typ položek, které nejsou shodné mezi modely. Dále zaměnit hodnoty některých výčtových typů [3]. Na druhé úrovni dochází k přejmenování sloupců. Oba modely sdílí nemalou množinu položek, které se liší pouze názvem a nikoli sémantickým významem. Při převodu těchto položek dojde pouze k přejmenování sloupců. A v poslední, třetí úrovni je provedena extrakce dat, které nejsou implicitně viditelné z modelu, ale je možné je odvodit na základě znalostí pracovního procesu. Například kurátor je zaznamenán u smlouvy, ale za současných podmínek se stává kurátorem přidruženého zdroje. Po proběhnutí tohoto skriptu, byla většina položek převedena do nového modelu a manuální zásahy jsou nutné především u tabulky kontaktů. Skript je možné nalézt v příloze D.

5.7.2 Dialog pro ruční převod

Druhým nástrojem, který si vyžádali uživatelé byl konverzní dialog. Tento formulář obsahuje dostatek editačních polí pro úpravu i vytvoření celého zdroje a závislých objektů. Je zde možné vytvářet nové smlouvy i vydavatele. Hlavní účel je právě editace nesprávně vyplněných zdrojů a tento účel úspěšně plní.

---

3. Například v sloupci doplněk zaměnit hodnotu „A“ za TRUE, formátovat telefonní čísla atd.

Závěr

Vytvoření tohoto software je velice důležité pro kurátory, kteří v Národní knihovně provádějí výběr dokumentů určených pro archivaci. Systém jim usnadňuje práci a zároveň kontroluje případné chyby, které mohou vzniknout lidským faktorem. Kurátor po příchodu na pracoviště získá okamžitý přehled o akcích, které vyžadují objekty, které má ve správě.

V práci jsem se zaměřil na podrobnou specifikaci, která je podkladem i pro samotné kurátory, takže ji mohou využít při zaučování nových pracovníků. Při návrhu jsem důkladně promyslel všechny známé situace v pracovním procesu a zapracoval je do datového modelu, aby byla flexibilita systému co nejvyšší. Aplikace je zaměřena především na komfort uživatelů a na tento fakt byl kladen důraz při implementaci systému.

Protože systém je určen pro reálné nasazení do procesu, je k dispozici zpětná vazba od jeho uživatelů, takže úpravy na systému mohou probíhat velice rychle a účelně. V současné chvíli kurátoři hodnotí mnou vytvořený systém kladně a sami nabízejí možnosti, jak jej dále rozšiřovat. Protože jsem poměrně do hloubky zpracoval specifikaci obsaženou v příloze, je možné systém doplnit nezávislými moduly, které rošřují funkcionalitu.

Velkým přínosem by bylo napojení systému na proces samotného sklízení dat, jak tomu je například u systému NetarchiveSuite. Datový model poskytuje možnost definovat četnost sklízení a tak je v budoucnu možné připojit sklízecího robota, který bude pracovat plně automaticky.

Dále by bylo vhodné více propracovat zasílaní oslovení vydavatelům, které mají charakter hromadné korespondence a tak je může poslat samotný systém. S tím souvisí i možnost vyplnění elektronické smlouvy o zdroji, která je podle posledních právních předpisů možná a tak nenutí vydavatele zasílat papírovou smlouvu poštou. Zapracování této funkcionality do systému by v mnohém usnadnilo úsilí kurátorů i vydavatelů.

Práce na projektu byla velice zajímavá a pro mne přínosná. Předpokládám, že touto problematikou se budu nadále zabývat a systém rozšiřovat i mimo tuto práci.

Literatura

[1] Archive.org. Available from World Wide Web: http://www.archive.org.

[2] AutoContractMarker modul. Available from World Wide Web: https://intranet.webarchiv.cz/wiki/index.php/AutoContractMarker.

[3] Creative Commons. Available from World Wide Web: http://creativecommons.org.

[4] Dublic Core Metadata. Available from World Wide Web: http://dublincore.org.

[5] Heritrix. Available from World Wide Web: http://crawler.archive.org.

[6] InnoDB. Available from World Wide Web: http://dev.mysql.com/doc/refman/5.0/en/innodb-overview.html.

[7] International Internet Preservation Consortium. Available from World Wide Web: http://www.netpreserve.org.

[8] KohanaPHP. Available from World Wide Web: http://kohanaphp.com.

[9] NetarchiveSuite. Available from World Wide Web: http://netarchive.dk/suite.

[10] PHP. Available from World Wide Web: http://www.php.net.

[11] Selekční kritéria projektu WebArchiv. Available from World Wide Web: http://www.webarchiv.cz/kriteria/.

[12] Unified Modeling Language. Available from World Wide Web: http://www.uml.org.

[13] Wayback. Available from World Wide Web: http://archive-access.sourceforge.net/projects/wayback/.

[14] Web Curator Tool. Available from World Wide Web: http://webcurator.sourceforge.net.

[15] Web Curator Tool User manual 1.4.0. Available from World Wide Web: http://webcurator.sourceforge.net/docs/1.4/wct-1.4.0-manual.pdf. s. 74.

[16] WebArchiv. Available from World Wide Web: http://www.webarchiv.cz.

[17] A. Brokeš. Projekt webarchiv - archiv českého webu. *Zpravodaj ÚVT MU*, roč. XVIII(4):s. 10–13, 2008. Available from World Wide Web: http://www.ics.muni.cz/zpravodaj/articles/578.html.

[18] P. Daněk and M. Rozehnal. Velký test PHP frameworků (2. díl). 2008. Available from World Wide Web: http://www.root.cz/clanky/velky-test-php-frameworku-2-dil/.

[19] A. Rauber and A. Aschenbrenner. Part of Our Culture is Born Digital - On Efforts to Preserve it for Future Generations. 2001. Available from World Wide Web: http://www.ifs.tuwien.ac.at/ifs/research/pub_html/rau_trans01/rau_trans01.html.

[20] P. Žabička. NEDLIB Harvester. 2000, roč. 4, č. 10. Available from World Wide Web: http://www.ikaros.cz/node/672. [cit. 2009-01-03].

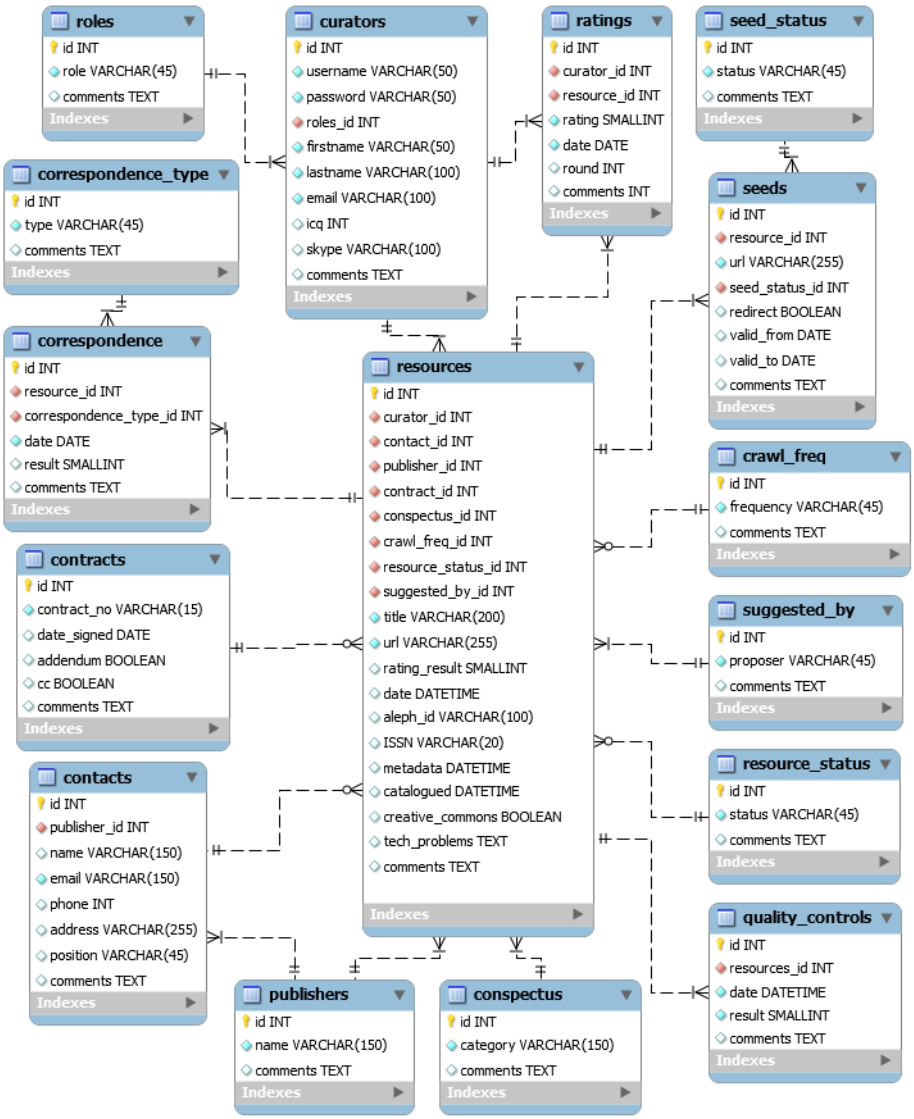
Dodatek A

Specifikace systému

Z důvodu rozsahu materiálu (15 stran) jsem specifikaci nevkládal do práce, ale vložil na přiložené CD. Jedná se o soubor s názvem *specifikace.pdf*.

Dodatek B

Datový model



Dodatek C

Ukázka rozhraní

Dodatek D

Migrační SQL skript

```
# Doplnek u smluv je treba prevest na binarni hodnotu
update smlouvy set doplnek = 1 where doplnek = 'A';
update smlouvy set doplnek = 0 where doplnek = 'N';

# Vlozit vydavatele
INSERT INTO publishers (id, name)
SELECT id, jmeno_vyd FROM vydavatele;

# Vlozit kontakty
INSERT INTO contacts (id, publisher_id, name, email, phone)
SELECT id, id, kont_osoba, email, tel FROM vydavatele;

UPDATE contacts SET phone = NULL WHERE phone = 0;

INSERT INTO conspectus(id, category) VALUES (1, 'unsorted');
INSERT INTO curators (id, username, password)
VALUES(1, 'unknown', MD5('password'));
INSERT INTO curators (username, password)
SELECT DISTINCT kdo, MD5('password') FROM smlouvy;

# Vlozime zdroje se smlouvou a kuratora urcime podle smlouvy
INSERT into resources
(id, title, url, ISSN, publisher_id, contract_id,
comments, conspectus_id, suggested_by_id, curator_id)
SELECT z.id, z.nazev, url, issn, z.vydavatele_id, smlouvy_id,
poznamka, 1, 5, c.id
FROM zdroje z, curators c, smlouvy s
WHERE c.username = s.kdo AND s.id = z.smlouvy_id;

# Vlozime zdroje bez smlouvy
INSERT into resources
(id, title, url, ISSN, publisher_id, contract_id,
comments, conspectus_id, suggested_by_id, curator_id)
SELECT z.id, z.nazev, url, issn, z.vydavatele_id, NULL,
poznamka, 1, 5, 1
FROM zdroje z WHERE smlouvy_id = 0;

# Pokud ma zdroj metadata, vlozime soucasne datum
UPDATE resources SET metadata = now()
WHERE id in (SELECT id from zdroje where metadata = 'A');
```

Dodatek E

Popis přiloženého CD

Přiložené CD obsahuje následující data:

- *wadmin.zip* – komprimovaný adresář se systémem WA Admin
- *wadmin.sql* – SQL skript pro vytvoření databáze
- *images/* – použité diagramy a jiné obrazové podklady
- *specifikace.pdf* – specifikace systému
- *thesis.pdf* – text práce