VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Úložiště digitálních dat pro potřeby ÚK VŠB-TU Ostrava I

Diplomová práce

2006 Bc. Dušan Jalůvka

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 4. května 2006 .....

Rád bych na tomto místě poděkoval všem, kteří mi s prací pomohli, protože bez nich by tato práce nevznikla.

Abstrakt

Tématem práce je nalézt nejlepší řešení pro implementaci a zprovoznění digitálního úložiště pro Ústřední knihovnu Vysoké školy báňské – Technické univerzity Ostrava. Práce se zabývá obecnými digitálními knihovnami a srovnává systémy DSpace a Eprints, používané jako digitální repozitáře. Z těchto dvou systémů se jeví jako lepší volba DSpace. Tento systém je podrobně popsán, stejně tak postup, jak byl systém přizpůsobován potřebám univerzitní knihovny. Popisuje také postup při převodu dat ze systému T-Series do DSpace. Práce obsahuje také příručku administrátora a knihovníka, která byla k systému DSpace vytvořena.

Klíčová slova: digitální knihovna, DSpace, repozitář, metadata, identifikátor

Abstract

The main subject of this work is to find the best solution for implementing digital repository at Central library in VSB – Technical university of Ostrava. This thesis concern with common digital library and compare two systems – DSpace and Eprints used as digital repository. DSpace appears as best choice between these systems. This system is described in detail and next is described building this repository in Central library. There are described problems which appear during building this repository system. Also there is described process to convert data from T-Series to DSpace. In the end is described user–administrator guidebook, which was created for DSpace.

Keywords: digital library, DSpace, repository, metadata, identifier

Seznam použitých zkratek a symbolů

J2SE	–	Java 2 Standard Edition
OAI	–	Open Archives Iniciative
PHM	–	Protocol for Metadata Harvesting
GPL	–	GNU Public Licence
DC	–	Dublin Core
MIT	–	Massachussetts Institute of Technology
HP	–	Hewlett Packard
CNRI	–	Corporation for National Research Initiative
PDF	–	Portable Document Format
RDF	–	Resource Description Framework
XML	–	eXtensible Markup Language
HTML	–	HyperText Markup Language

Obsah

Seznam tabulek

Seznam obrázků

1 Úvod

V dnešní moderní počítačové době stoupá potřeba uchovávat a zpřístupňovat dokumenty elektronickou formou. Šíření informací uložených klasicky v tištěné podobě má svá omezení a v éře globálně používaného média, Internetu, jde o překonanou praxi. Z tohoto pohledu nabízí digitální repozitáře mnoho podob využití.

Široké uplatnění nacházejí digitální repozitáře v knihovnách a výzkumných centrech. Klasické knihovny zřizují tzv. digitální knihovny pro získávání, uchovávání, zpřístupňování a organizování dokumentů, které mají k dispozici v elektronické podobě. Výzkumná centra využívají digitální úložiště pro sdílení a prezentování výsledků výzkumů a jiných činností.

Oproti klasické knihovně přináší digitální knihovna řadu výhod. Především dovede uchovat daleko větší množství informací a dokáže tyto dokumenty uchovávat ve stále stejné kvalitě. Další výhodou digitálních úložišt a knihoven je fakt, že nemusí uchovávat jen textové dokumenty, ale dokáže uchovat různé formáty dat od fotografií technických výkresů, až po videozáznamy různých konferencí a podobně. Takové výhody sice mohou poskytnout i jiné technologie než digitální knihovny, ale ty zase mají jiné nevýhody, které znemožňují použití v plném rozsahu. Tyto nevýhody budou popsány dále v textu.

Cílem práce by mělo být poskytnutí digitálního úložiště Ústřední knihovně Vysoké školy báňské – Technické univerzity Ostrava, která by pomocí něho zpřístupňovala digitální verze naskenovaných článků ze sborníků vědeckých prací a elektronické verze kvalifikačních prací studentů. Mělo by proběhnout porovnání několika připravených systémů pro digitální úložiště a na základě zkušeností s nimi by měl být vybrán jeden z nich. Do tohoto systému by měly být převedeny data z jiných systémů, které jsou pro tyto potřeby nevyhovující.

Popis kapitol

V této práci se budu zbývat problematikou:

1. První kapitola je úvod,
2. bude podrobně rozvedena problematika digitálních knihoven, budou popsány jejich výhody a nevýhody ve srovnání s klasickými knihovnami a používané technologie a standardy,
3. bude popsán proces výběru správného systému pro nasazení a v rámci tohoto procesu také srovnání dvou rozšířených systémů DSpace a Eprints. Vybraný systém bude důkladněji popsán,
4. v Ústřední knihovně Vysoké školy Báňské – Technické univerzity Ostrava jsem vyzkoušel uvést do provozu software, který byl vytvořen právě za účelem uchovávání digitálních dat. Provedl jsem na něm řadu úprav, aby přesně splňoval požadavky kladené na uchovávání digitalizovaných článků a vysokoškolských kvalifikačních prací. V této kapitole popíši postup při úpravách a implementaci tohoto digitálního repozitáře,

5. kapitola 5 popisuje příručku pro administrátora a knihovníka, která byla v rámci diplomové práce vytvořena. Měla by ulehčit pochopení složitějších úkonů při administraci systému.

6. v závěru se pokusím zhodnotit výsledky práce a zmíním se o tom, jak by se dal systém dále rozšířit.

2 Teorie digitálních knihoven

Nyní bude podrobně popsána problematika digitálních knihoven, co taková digitální knihovna znamená a jak se liší od klasické knihovny. Popíši výhody a nevýhody digitálních knihoven a ve stručnosti se zmíním o technologiích spojených s digitálními knihovnami a úložištěmi.

2.1 Co je digitální knihovna

Digitální knihovna je knihovna, ve které největší zastoupení zdrojů představují dokumenty v elektronické podobě. Dokument v elektronické podobě může být vytvořen přímo pomocí počítače nebo se pomocí nějakých elektronických zařízení (například skeneru) převedou již existující tištěné dokumenty do podoby digitálních dat. Tato data mohou být pomocí počítače převedena do čitelné podoby, například vytisknuty na papír nebo zobrazeny na monitoru počítače. Takový digitální dokument pak může být uložen na jednom místě a může se k němu přistupovat pomocí počítačové sítě nebo přímo lokálně.

Definice digitální knihovny by mohla znít například takto: „Digitální knihovna je sbírka digitálních objektů, obsahujících text, video, zvuk, technické zprávy, . . . “.

Společným prvkem digitálních a klasických „kamenných“ knihoven je text. Textové objekty v digitálních knihovnách nazýváme *dokumenty*. Digitální knihovny ale nemusí obsahovat jen textové dokumenty, jejich univerzálnost dovoluje uchovávat například hudbu v mnoha reprezentacích (audio nahrávky, notové zápisy, MIDI nahrávky a další), technické dokumenty (výkresy v digitální podobě, specifikace, a další), video (záznamy přednášek, snímky ze stacionární družice, a další), software, obrázek (umělecké fotografie, geografické mapy, a další), a jiné. Protože by takové typy digitálních objektů byly špatně vyhledatelné, ke každému takovému objektu se v digitálních knihovnách uchovávají tzv. *metadata*.

Metadata jsou popisná data dokumentu (data o datech), která popisují obsah digitálních objektů pro jejich snadnější vyhledávání. Metadata mohou obsahovat informace o autorovi, formátu, velikosti, stáří a další údaje o digitálním objektu v archivu. Podrobnější popis metadat bude v kapitole 2.5.1.

Hlavními funkcemi digitálních knihoven je získávání, zpracovávání, uchovávání, zpřístupňování a ochrana dat. Získáváním dat se rozumí neustálé přidávání nových objektů do archivu digitální knihovny. Přidávání může provádět například knihovník ručním vkládáním nových objektů a vypisováním metadat nebo automatizované mechanismy pro sběr dat. Takovým mechanismem je například protokol OAI-PHM, který bude podrobněji popsán v kapitole 2.5.2. Zpracováním je myšleno správné třídění digitálních objektů podle jejich obsahu a přiřazení jednoznačného identifikátoru, tzn. přidat například digitální mapu povrchu Marsu do sbírky digitálních map a přiřazení správného a jedinečného identifikátoru. Identifikátorům se podrobněji věnuje kapitola 2.5.3. Uchováváním rozumíme uložení digitálního objektu do společného archivu všech dat a udržování dat v takovém formátu, aby byly kdykoliv čitelné (tzn. aby byly v takovém formátu, že půjdou otevřít běžnou aplikací). Metody takového udržování formátu jsou podrobněji popsány v kapitole 3.4.7. Součástí uchovávání je i *ochrana dat*, která by

měla zajistit bezpečné uložení dat a pravidelné zálohování celého obsahu digitálního archivu.*Zpřístupňováním* rozumíme zajištění dohledatelnosti objektu, aby se podle zadaných metadat bylo možné dostat ke konkrétnímu digitálnímu objektu.

2.2 Pohled do historie knihoven

První zmínka o „digitální knihovně“ je z roku 1945, když se Vannevar Bush ve svém vizionářském článku *As We May Think* [7] zabýval efektivnějším „automatizovaném“ zpracování odborných informací. Za dalšího „průkopníka“ digitálních knihoven bývá považován J. C. R. Licklider, který v roce 1965 publikoval knihu *Libraries of the future* [8], v níž popisuje výzkum a vývoj potřebný k realizaci opravdové digitální knihovny.

První skutečné digitální knihovny se začaly objevovat s větším rozšířením výpočetní techniky a počítačových sítí počátkem 90. let. Problematika digitálních knihoven je natolik obsáhlá, že vývoj probíhá neustále a pořád se objevují nové myšlenky na vylepšení stávajících systémů. Více se o historii digitálních knihovnách můžete dočíst v [6].

2.3 Výhody a nevýhody digitálních knihoven

Tak jako mají digitální knihovny výhody oproti klasickým knihovnám, mohou z některých výhod plynout nevýhody, které klasické knihovny nemají. Rozdíly mezi oběma typy knihoven lze shrnout takto.

Nejpodstatnějšími výhodami jsou:

- proti klasickým knihovnám nejsou digitální knihovny omezeny skladovacím prostorem, protože digitalizovaná data zabírají zanedbatelný prostor. Pro srovnání: například jedna rozsáhlá klasická knihovna s rozlohou jedné budovy by mohla fungovat v rámci jedné místnosti i s obsluhou,
- do digitální knihovny nemusíte chodit, takže digitální knihovnu může navštívit současně daleko více lidí. Navíc můžete „navštívit“ digitální knihovnu vzdálenou od vás několik tisíc kilometrů rychleji, než místní klasickou knihovnu,
- s tím také souvisí provozní doba knihovny, kterou můžete navštívit v kteroukoliv denní či noční dobu,
- stejné dokumenty může číst více lidí najednou,
- v digitální knihovně můžete vyhledávat podle fráze nebo klíčových slov současně v celém obsahu digitální knihovny,
- digitální materiál se daleko lépe uchovává a kopíruje, takže kopie originálu neztrácí na kvalitě. Navíc může být obsah dokumentu konvertován do modernějších formátů, které se časem vyvinou,
- díky propojení všech digitálních knihoven internetem může knihovna nabídnout při hledání i dokumenty, které jsou fyzicky uloženy v jiné vzdálené digitální knihovně,

- cena provozu digitální knihovny se může zdát nižší než u klasické knihovny, která musí platit zaměstnance, a jiné poplatky. U digitální knihovny se musí investovat nemalé prostředky do převodu materiálů do digitální podoby a do zajištění online přístupu (internetové připojení, pořízení a provoz serverů). Takže se u menší knihovny může při pořízení digitální verze knihovny cena jevit jako nevýhoda.

Nevýhod digitálních knihoven není tolik jako výhod, ale pro někoho mohou být docela zásadní. Patří mezi ně:

- na mnoho dokumentů se vztahují autorské práva, takže nemohou být volně přístupné všem. Proto obsahem digitálních knihoven bývají většinou veřejně dostupné dokumenty nebo dokumenty vlastní produkce,
- někteří lidé tvrdí, že tištěné dokumenty se čtou mnohem lépe než text zobrazený na monitoru, ale to může záviset na prezentaci textu a preferencích čtenářů. Digitální knihovna také nemůže nahradit prostředí a atmosféru klasické knihovny.

2.4 Možnosti využití digitálních knihoven

Digitální knihovny mohou mít daleko větší uplatnění, než pouhé nahrazení nebo vylepšení stávajících klasických knihoven. Používají se také jako mohutná skladiště pro ukládání dokumentů nejrůznějšího typu v digitální podobě.

2.4.1 Institucionální repozitáře

Institucionální repozitář slouží ke sběru digitálního materiálu z výzkumných laboratoří, akademických výzkumných projektů, elektronických kvalifikačních prací a dalších. Obsah těchto typů dokumentů zpřístupňuje akademické komunitě a částečně i široké veřejnosti. Vybudování takového akademického repozitáře bylo hlavním cílem této práce. Postup při budování institucionálního repozitáře je podrobněji popsán v kapitole 3.5.

2.4.2 Digitální archivy

Digitální archivy se od digitálních knihoven liší v několika aspektech. Převážně obsahují dokumenty z vlastních zdrojů jako jsou dopisy a další dokumenty vytvořené institucí, místo dokumentů jako jsou knihy nebo časopisy.

Dokumenty uložené v digitálním archivu mají jedinečný obsah a nebývají vkládány do jiných archivů. To znamená, že je nemůžeme najít nikde jinde, než v konkrétním digitálním archivu, narozdíl od knih v knihovnách.

Obsah digitálního archivu většinou bývá řazen do skupin dokumentů, kdežto knihy v knihovnách bývají řazeny jako jednotlivé položky. Dokumenty mohou být například řazeny podle původu vzniku (tvůrce nebo organizace) a jedinečného pořadí (datum vytvoření).

S digitálními archivy souvisí také řízený koloběh dokumentů po instituci nebo firmě. Ten je podrobněji popsán v kapitole 3.4.4.

V dalším textu bude pod pojmem digitální knihovna zahrnut i digitální archiv a institucionální repozitář.

2.5 Technologie a standardy používané v digitálních knihovnách

Při vývoji digitálních knihoven a repozitářů se vyvinula řada standardů používaných pro fungování digitálních knihoven. Patří mezi ně standardy pro popis digitálních objektů, pro interakci mezi jednotlivými digitálními knihovnami na internetu nebo také standardy pro jednoznačné identifikátory. V této části budou popsány standardy, se kterými jsem pracoval v průběhu řešení práce a které budou dále v textu používány.

2.5.1 Metadata

Metadata se dají přirovnat ke katalogovým lístkům v knihovně, které obsahují informace o knize a jejím umístění. Metadata jsou informace o digitálních objektech („Data about data“), které určují, jak jednoznačně popsat atributy digitálních záznamů. Můžeme je rozdělit podle možností použití na metadata *popisná*, *strukturální* a *administrativní*. *Popisná* metadata slouží k popisu digitálního objektu, aby jej bylo možné vyhledat a identifikovat. *Strukturální* metadata určují formát objektu, strukturu souboru nebo jeho velikost. *Administrativní* metadata se používají pro řízení autentikace přístupu a zálohování. Většinou nejsou uloženy s popisnými metadaty, ale ukládají se s nebo do digitálních objektů. Při vývoji metadat se vyvinula celá řada standardů pro zápis metadat digitálních objektů.

Známým standardem pro popis metadat v prostředí webu a digitálních dat je *Dublin Core* (používá se zkratka DC) [12]. První verze DC vznikla v březnu roku 1995 na semináři, kterého se účastnilo 52 specialistů z oborů knihovnictví, zpracovávání textů, počítačoví experti a další. Původně měl Dublin Core sloužit jako popis zdrojů sestavený přímo autorem, ale nakonec zaujal instituce jako muzea, knihovny a další.

Dublin Core definuje patnáct základních elementů pro popis dat a ke každému elementu ještě několik zpřesňujících prvků, tzv. kvalifikátorů. Dohromady tvoří dobrou základnu pro popis dat různých typů dokumentů. Všechny elementy jsou nepovinné a mohou se opakovat. Základních 15 elementů popisuje tabulka 1. Každý kvalifikovaný element má dáno schéma zápisu, aby byla dodržena jednotnost v zápisu. Například jazyk se zapisuje dle normy ISO 639 (cs, en, sk, . . .) nebo datum dle normy ISO 8601 (2006-04-23, 2006-04 nebo jen 2006). Seznam všech elementů, kvalifikátorů a jejich podrobný popis můžete nalézt v [12].

Zatímco Dublin Core určuje popisná a částečně strukturální metadata, standard METS (Metadata Encoding & Transmission Standard) definuje pomoci jazyka XML popisná, strukturální i administrativní metadata. Jazyku XML je věnována kapitola 4.1.7. METS definuje sedm sekcí pro popis dat:

1. hlavičku METS, popisující metadata samotné,
2. popisná metadata,
3. administrativní metadata,

Jméno elementu	Význam elementu
Title	Název
Creator	Tvůrce
Subject	Předmět nebo klíčová slova
Description	Popis
Publisher	Vydavatel
Contributor	Přispěvatel (autor, spoluautor, sponzor, . . .)
Date	Datum (vydání, vytvoření, zpřístupnění, . . .)
Type	Typ práce (článek, obrázek, zvuk, . . .)
Format	Formát dokumentu (počet stran, datový formát, . . .)
Identifier	Identifikátory (ISSN, ISBN, handle, . . .)
Source	Zdroj dokumentu
Language	Jazyk dokumentu
Relation	Vztah (je částí, je založen, . . .)
Coverage	Rozsah (prostorový, časový)
Rights	Práva (autorská, distribuční)

Tabulka 1: Základní elementy Dublin Core

4. sekce souborů, obsahuje seznam všech souborů,
5. strukturální data, informace, kam zařadit dokument,
6. strukturální odkazy, dovolují provázat data v hierarchii,
7. chování, určuje, jak zacházet s daty.

Jako rozšíření METS byl vyvinut standard MODS (Metadata Object Description Schema. Podrobnější popis viz [13, 14].

Posledním zástupcem z řady metadat, o kterém se tady zmíním je standard RDF (Resource Description Framework). Za pomoci jazyka XML definuje, jak popsat metadata (například Dublin Core). RDF našel uplatnění především v oblasti sémantického webu. Více informací viz [15].

2.5.2 Interoperabilita

Požadavkem správce knihovny je, aby si různé digitální knihovny (repozitáře, archivy) uměly vyměňovat metadata jednotným způsobem a aby v rámci jedné digitální knihovny byl vyhledatelný alespoň částečný obsah jiné digitální knihovny. Dalším problémem je provázání takto získaných metadat se skutečnými daty v libovolné jiné digitální knihovně. Takové chování může zajistit jedině jednotný protokol pro interoperabilitu mezi knihovnami. Takových protokolů bylo navrženo a implementováno několik, jako příklad bych uvedl OAI–PMH [16], Z39.50 [17], OpenURL [18] a další. Protože v rámci mé práce jsem se setkal s protokolem OAI–PMH, podrobněji se zmíním o tomto protokolu.

Open Archives Initiative Protocol for Metadata Harvesting (dále jen OAI–PMH) je jednoduchý protokol pro získávání metadat z jiných digitálních knihoven. První verze protokolu vznikla počátkem roku 2001 a požaduje, aby metadatadovým standardem pro vzájemnou komunikaci byl minimálně Dublin Core. Je založen na komunikačním protokolu HTTP a formátu XML.

Protokol definuje 5 základních pojmů:

- *Resource* – entita, která je popisována daným metadatovým objektem,
- *Repository* – repozitář, který je provozovaný poskytovatelem dat a prostřednictvím protokolu OAI–PMH poskytuje metadatové záznamy,
- *Harvester* – klientský program, provozovaný poskytovatelem služeb za účelem získání metadat z repozitářů,
- *Item* – metadatový objekt v repozitáři, ze kterého se odvozují metadatové záznamy pro protokol,
- *Record* – metadatový záznam v konkrétním formátu, získaný nebo odvozený z daného metadatového objektu.

Díky tomuto rozdělení umožňuje protokol vytvářet informační centrály a poskytovat další služby, které dovolují vyhledávat pomocí metadat digitální objekty uložené v různých digitálních knihovnách. Více informací o protokolu OAI–PMH naleznete na [16, 9].

2.5.3 Identifikátory

V klasických knihovnách se používají identifikátory jako ISSN, ISBN, ISTC a další, stejně tak se pro digitální objekty používají zvláštní identifikátory v prostředí internetu. Na identifikátory je kladeno několik požadavků:

- *jednoznačnost* – identifikátor by měl být jedinečný v celé globální síti internetu,
- *perzistence* – identifikátor by měl být objektu přiřazen natrvalo, ikdyž bude objekt zrušen,
- *škálovatelnost* – jméno by mělo být použitelné pro jakýkoliv možný typ entity.

Obecným schématem pro identifikátory se stal Uniform Resource Name (URN), který definuje identifikaci objektu nezávisle na jeho umístění. Nalezení správného umístění dle tohoto identifikátoru by měl řešit jistý směrovací protokol. V praxi se jedná většinou o servery, které na dotaz s identifikátorem vrací jeho fyzické umístění na síti.

Obecný identifikátor by měl být ve tvaru „urn:nid:nss“, kde *nid* (namespace identifier) je řetězec identifikující použitý identifikační systém (například doi, hdl, issn a další) a *nss* (name specific string) je řetězec s konkrétním identifikátorem. Konkrétní identifikátor patřící knize [1] by byl *urn:isbn:8072269194*.

Jeden z konkrétnějších identifikačních schémat je CNRI handles [19]. Schéma je založeno na protokolu HTTP a je plně kompatibilní s konceptem URN. Umožňuje přiřazování,

správu a rozpoznávání perzistentních identifikátorů digitálních objektů. Identifikátor má tvar „hdl:cnri.dlib/locid“, kde první část (nazývaná prefix) *cnri.dlib* definuje pojmenovávací autoritu. První část se přiděluje globálně, druhá část lokálně v rámci instituce. Část *locid* je jakýkoliv řetězec jedinečný v rámci pojmenovávací autority. Identifikační schéma CNRI handles používá i systém DSpace, který byl využit v prostředí Ústřední knihovny, a který je popisován dále.

Dalším z identifikačních schémat je Digital Object Identifier (DOI) [20]. DOI používá tvar jako CNRI handle, ale v části pro pojmenovávací autoritu je konstanta 10 pro odlišení od ostatních handle systémů. Narozdíl od CNRI handles, který je zdarma, je schéma DOI zpoplatněno. Platí se za samotnou registraci organizace, i za přidělení čísla DOI.

2.6 Shrnutí

Téma digitálních knihoven a shromažďování a zpřístupnění informací jejich prostřednictvím otvírá prostor pro další zdokonalení, přestože již je k dispozici řada dílčích standardů a protokolů, které umožňují fungování rozsáhlých digitálních knihoven. Mnoho problémů je stále ještě otevřených a na jejich řešení se neustále pracuje.

Nedořešena je otázka přístupu k uloženým objektům, otázka čtení a kopírování plných textů neoprávněnými osobami, kde je potřeba zajistit řízení přístupu k uloženým objektům a nedovolit čtení nebo kopírování neautorizovaným osobám. U klasických knihoven lze tento problém řešit omezením fyzického přístupu k dokumentům.

Další část práce se zabývá již samotnou implementací digitálního repozitáře v prostředí Ústřední knihovny vysoké školy báňské – Technické univerzity Ostrava. V případě zájmu o teorii digitálních knihoven se můžete obrátit na dříve citované dokumenty nebo použít některou digitální knihovnu (například [21]) a zkusit najít požadované informace tam.

3 Analýza řešení problému

Tato kapitola se věnuje analýze problému, popisuje postup při rozhodování pro vhodný systém digitálního repozitáře a popisuje vybraný systém DSpace. Analýza by měla jednoznačně definovat jednotlivé požadavky na systém. Nakonec je popsán obecný postup při implementaci digitálního repozitáře.

Hlavním cílem práce je poskytnutí digitálního úložiště dat, do kterého by Ústřední knihovna VŠB–TU Ostrava mohla vkládat elektronické verze dokumentů. V první fázi se předpokládá uložení vysokoškolských kvalifikačních prací a sborníků vydávaných Vysokou školou báňskou – Technickou univerzitou Ostrava, které se knihovna rozhodla zpřístupnit v elektronické verzi. Předpokládá se, že převod sborníků do elektronické verze potrvá velmi dlouho (řádově několik let), protože je třeba naskenovat více než 3300 článků. Asi 800 článků je již v elektronické podobě a je potřeba tyto články již zahrnout do digitálního úložiště společně s jejich metadaty, která budou importována jednorázově. K těmto metadatům musí systém později umožnit naskenovaný plný text. Systém by měl také umožnit vložit metadata o vysokoškolských kvalifikačních pracích, které nejsou k dispozici v elektronické verzi, ale jsou k dispozici v tištěné podobě. V další fázi by měl systém sloužit jako úložiště pro digitalizované staré tisky (do roku 1800) z archivního fondu knihovny a mělo by se připravit univerzální rozhraní pro vkládání různých výzkumných prací a výsledků vědecké činnosti univerzity.

Systém musí umožňovat fulltextové vyhledávání metadat a procházení jednotlivých záznamů podle různých kritérií (autor, název, datum).

Dalším požadavkem je, aby přístup k plným textům vysokoškolských kvalifikačních prací a článkům ze sborníků byl povolen jen studentům a zaměstnancům školy. Ostatním musí být zpřístupněny informace o všech záznamech, ale ne soubory s plnými texty. Plný přístup by tedy měli mít všichni ti, kteří se do systému přihlásí svým školním LDAP heslem.

Přístup a přihlašování uživatelů by mělo probíhat přes zabezpečený protokol HTTPS, aby nebylo možné na trase mezi uživatelem a serverem odposlechnout heslo a další citlivé informace.

Dílčím úkolem je také importovat část bibliografických dat o vysokoškolských kvalifikačních pracích a článcích ze sborníku z jiného systému, v němž byla data doposud uložena. Jedná se o knihovnický systém T-Series. Systém T-Series používá k exportu dat svůj vlastní formát souboru, takže bude třeba vytvořit program, který vyexportovaný soubor z T-Series převede do metadatového formátu Dublin Core zapsaného v XML.

3.1 Současný stav

V současné době je proces zpřístupnění vysokoškolských kvalifikačních prací velice zdlouhavý a nedokonalý. Práce se odevzdávají jednotlivým katedrám a po obhájení se všechny práce posílají do knihovny ke zpracování. Tam jsou kvalifikační práce pracovníky knihovny katalogizovány a údaje o práci jsou vloženy do systému T-Series. Fyzicky je práce přístupná ve studovně, kde je k dispozici k nahlédnutí. V systému T-Series je vedeno více než 20 tisíc prací.

Články ze sborníků Vysoké školy báňské – Technické univerzity Ostrava a bibliografické záznamy o článcích ze sborníků se vytvářejí a uchovávají v systému T-Series. Byl započat proces převodu článků do elektronické podoby. Z celkových asi 3300 článků je do elektronické verze převedeno asi 800 článků a jsou uloženy v PDF formátu.

3.2 Případy použití

Případ použití (use case) naznačuje hranice systému, základní poskytované funkce a typy uživatelů, kteří budou se systémem pracovat. Uživatelům jsou přidělovány role podle toho, jaké funkce systému smějí využívat. Případy použití jsou zachyceny v textové a grafické podobě. K diagramu je přiložen také textový popis jednotlivých rolí a případů užití. Pro větší přehlednost se diagramy mohou rozdělit hierarchicky do několika podrobnějších diagramů. Diagram na obrázku č. 1 ukazuje, jaké případy použití by měl systém umožňovat.

Uživatel v roli *administrátor* bude moci provádět veškeré operace se systémem, nastavovat systém, určovat, jaká oprávnění budou mít jednotliví uživatelé. Bude moci vytvářet nové kolekce dokumentů a přidělovat jim uživatele, kteří budou zodpovědní za obsah a správnost vyplněných metadat.

Role *správce kolekce* umožní uživateli v této roli provádět operace spojené se správou přidělené kolekce (editace, mazání, přidávání dokumentů). Tuto roli přiděluje uživateli jedině administrátor.

Role *přihlášený uživatel* a *anonymní uživatel* reprezentují návštěvníky digitálního úložiště, kteří mohou vyhledávat dokumenty, procházet dokumenty podle různých kritérií. Přihlášení uživatelé mohou navíc, v závislosti na nastavení kolekce, vkládat nové dokumenty do kolekcí a mohou mít také přístup k plným textům dokumentů uložených v úložišti.

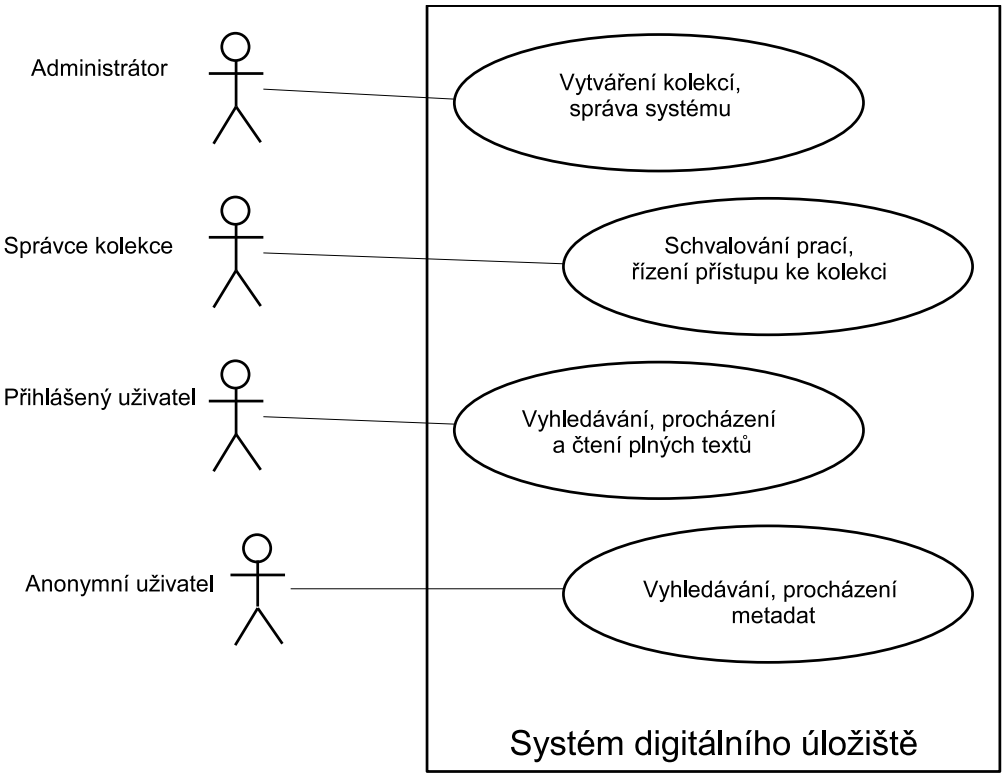
3.3 Porovnání systémů Eprints a DSpace

Jako vhodné řešení pro implementaci digitálního úložiště byly vybrány dva systémy DSpace a Eprints, které jsou používány univerzitami na celém světě. Z těchto dvou systémů bude po srovnání a zkušební instalaci vybrán vhodnější a bude použit jako oficiální digitální úložiště ústřední knihovny. Obecné srovnání těchto a dalších systémů pro zpřístupňování digitálních dokumentů můžete najít v [3].

Oba systémy jsou volně dostupné (open source), takže jsou k nim dispozici i zdrojové kódy a je možno je jakkoliv modifikovat. Já jsem v rámci své práce pracoval se systémem DSpace a kolega Michal Pastuszek ve své práci [5] testoval systém Eprints. Po vzájemných konzultacích a srovnáních se nakonec ukázalo, že pro potřeby Ústřední knihovny bude lépe vyhovovat systém DSpace, na kterém se provedou mírné úpravy.

K tomuto rozhodnutí nás vedly především tyto důvody:

- DSpace má jednodušší instalaci a méně požadavků na hostitelský systém,
- je snadněji modifikovatelný díky poskytnutému programátorskému rozhraní a předpřipravenými skripty,



Obrázek 1: Diagram případů použití

- díky definici vkládacích formulářů v jediném XML souboru má DSpace tyto formuláře jednodušeji modifikovatelné než Eprints a nejsou velké problémy s přidáním nových formulářů pro jiné typy kolekcí,
- DSpace poskytuje lepší možnosti pro vytvoření obecného repozitáře obsahujícího mnoho typů digitálních objektů,
- administrátorské rozhraní umožňuje provádět více operací než v Eprints,
- ze zkušeností knihovníků, které se vyskytly při testovacím provozu, vyplývá, že DSpace má více intuitivní ovládání než Eprints,
- a další drobnosti, které se projevili při administračních pracích se systémem DSpace.

3.4 Popis DSpace

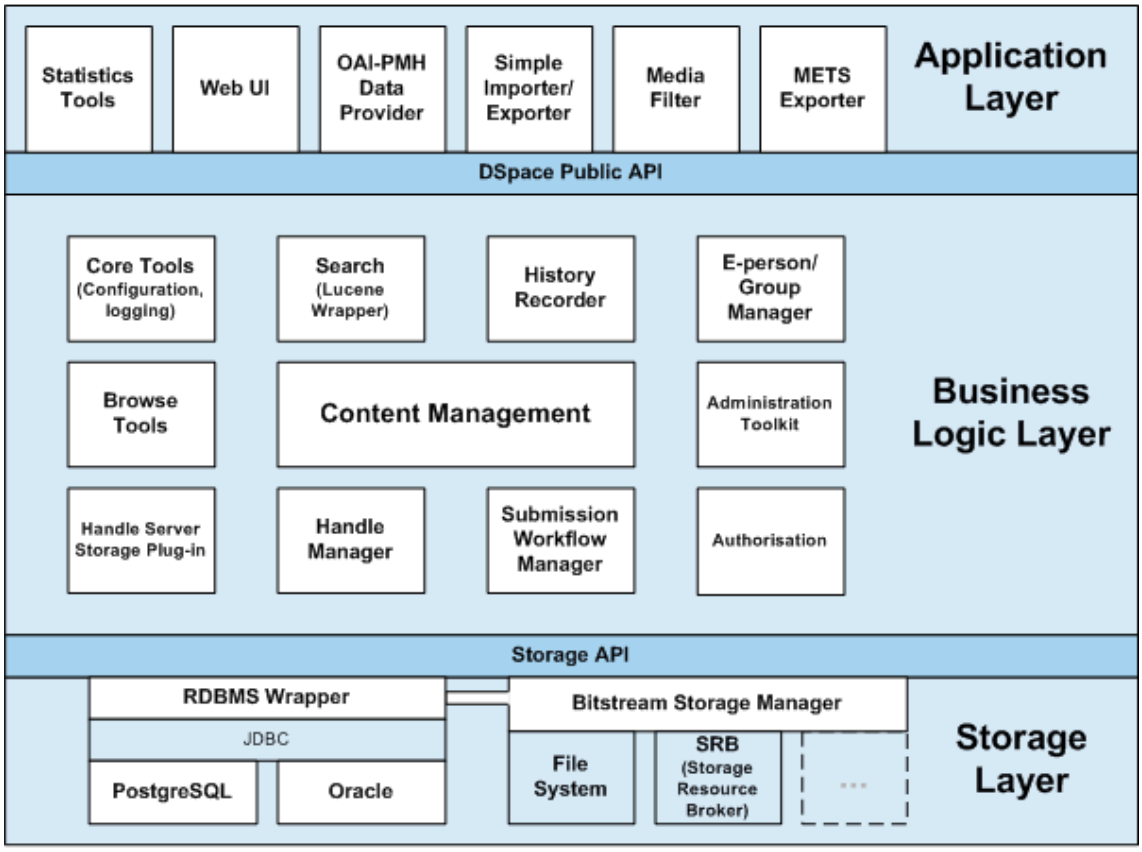
DSpace [11] vznikl v roce 2002 jako výsledek spolupráce Massachussetts Institute of Technology (MIT) a Hewlett Packard (HP). Hlavním cílem bylo vytvořit univerzální digitální repozitář pro potřeby MIT, ale také poskytnout univerzální základ pro použití v jiných institucích, především z akademicé sféry. Je distribuován pod licencí BSD (Berkeley Software Distribution), což je svobodná licence, která umožňuje jakékoliv změny v poskytnutých zdrojových kódech při dodržení jistých podmínek [22].

DSpace používá řadu technologií a standardů, aby poskytl co možná nejkomfortnější práci se systémem. Je použitelný pro uchovávání různých druhů archivovaných materiálů, jako jsou články, technické zprávy, elektronické kvalifikační práce, obrazová data, video data, audio data, výukové materiály a další. Mnoho organizací používá DSpace jako institucionální repozitář nebo digitální knihovnu.

Systém má v sobě implementovánu řadu nástrojů, které zjednodušují práci se systémem a jeho administraci. Umožňuje nastavení různých bezpečnostních politik a uživatelských účtů, definuje třístupňové schéma řízení koloběhu dokumentů, dohlížení vedoucích nad prácemi studentů, rozesílání emailů se seznamem nových příspěvků v archivu, import a export obsahu archivu do formátu XML a metadatového schématu Dublin Core a další. Některé z těchto nástrojů budou podrobněji popsány v následujícím textu, který se věnuje popisu DSpace verze 1.3.2.

3.4.1 Systémová architektura

Systém je naprogramován v jazyce Java v kombinaci s dynamicky generovanými HTML stránkami pomocí Java Server Pages (JSP). Původně byl napsán pro operační systém UNIXového typu (Linux, HP/UX, Solaris), ale jelikož jsou použity multiplatformní technologie, mělo by být možné nainstalovat DSpace i na jiné systémy (například MS Windows v kombinaci s Cygwin). Pro ukládání metadat je použita relační databáze PostrgreSQL nebo Oracle ve spojení s ovladačem JDBC (Java Dababase Connector). Jako servlet kontejnér může být použit server Apache Tomcat, Jetty nebo Caucho Resin. Doporučována je kombinace Linux, PostgreSQL a Tomcat.



Obrázek 2: Architektura DSpace

DSpace je postaven na třívrstvém modelu, který je znázorněn na obrázku 2 (obrázek převzat z [2]). Jednotlivé vrstvy spolu komunikují prostřednictvím společného rozhraní a zajišťují tak oddělení a nezávislost částí systému.

Nejnižší vrstva (Storage layer) zajišťuje ukládání dat a metadat prostřednictvím databáze a manažera dat (Bitstream Storage Manager). Manažer dat dovoluje uložení buď na lokální souborový systém, nebo na vzdálený server (podrobněji popsáno v kapitole 3.4.3).

Prostřední vrstva (Bussines Logic Layer) poskytuje rozhraní (Public API) pro veškeré funkce systému. Jak je vidět z obrázku 2, zajišťuje autorizaci uživatelů, vyhledávání a prohlížení digitálních záznamů, přidělování identifikátorů, schvalování a řízení koloběhu dokumentů a další správu obsahu digitálního úložiště.

Nejvyšší vrstva (Application Layer) poskytuje rozhraní pro styk úložiště s okolním světem. Pomocí volání funkcí z Public API dává k dispozici nástroje pro tvoření statistik, webové rozhraní pro uživatele a správce, nástroj pro import a export záznamů, rozhraní protokolu OAI-PMH pro získávání metadat z jiných digitálních úložišť, media filter pro tvorbu náhledů obrázků a jiného zpracování uložených souborů a také umožňuje exportovat obsah úložiště do metadat standardu METS.

Pro veškerá rozhraní je poskytnuta kvalitní dokumentace s nabízenými funkcemi, takže je možné kdykoliv doprogramovat vlastní verzi nebo změnit stávající část systému.

3.4.2 Datový model DSpace

Způsob jak DSpace zachází se záznamy a uživateli znázorňuje diagram datového modelu na obrázku 3 (obrázek převzat z [2]). DSpace nazývá veškerý digitální obsah v archívu jako *položky* (Items). Položka se skládá ze všech souborů s digitálním obsahem a z popisných metadat uložených v Dublin Core.

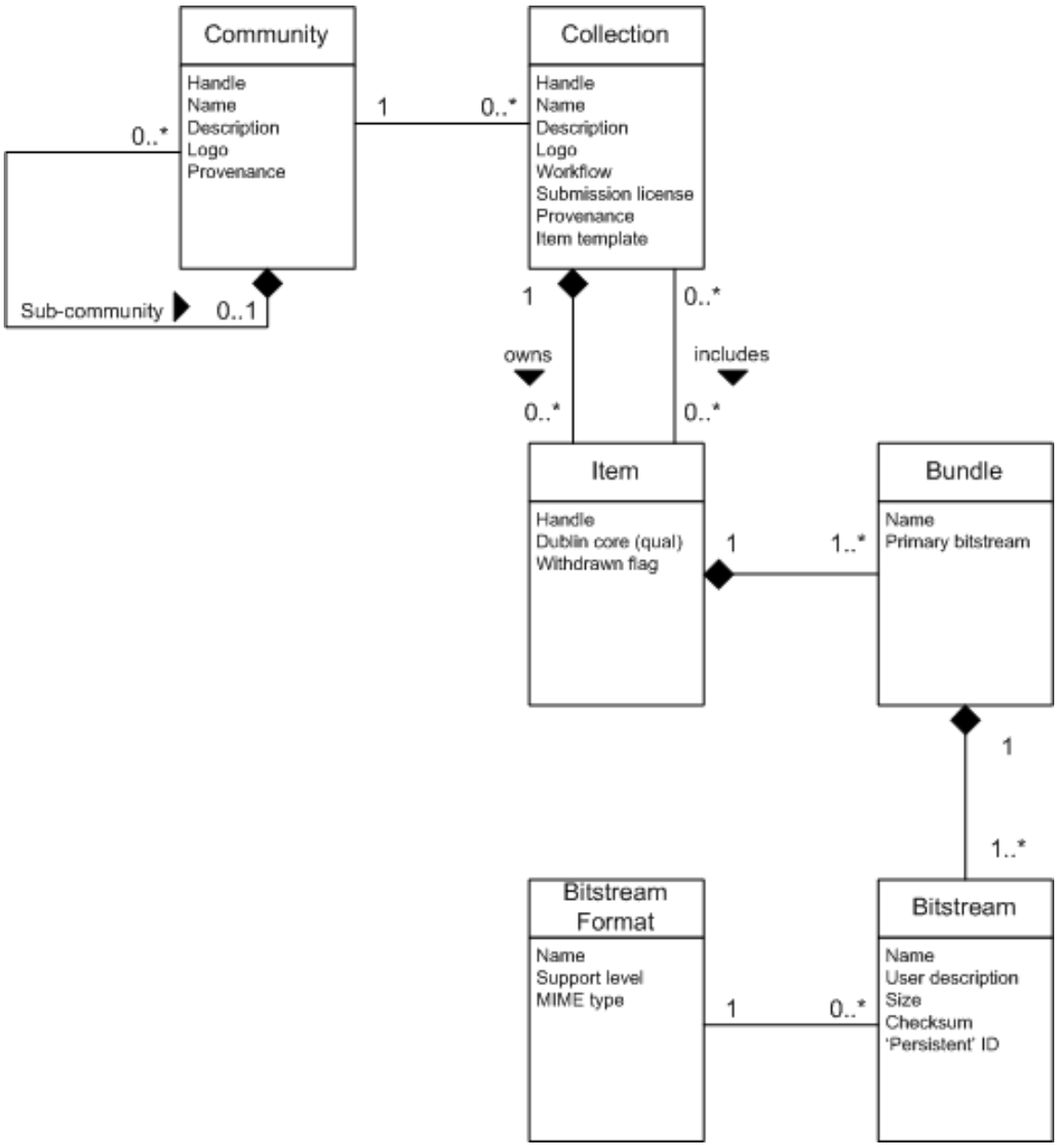
Hlavním prvkem je komunita. Komunita představuje organizační jednotky nebo skupiny uživatelů, kteří se rozhodli ukládat do DSpace své položky. V univerzitním prostředí to většinou bývají fakulty, katedry nebo výzkumná pracoviště. Každá komunita může být rozdělena do několika podkomunit.

Aby si komunity mohli ukládat v DSpace položky různého charakteru, každá komunita si zakládá kolekce položek. Například kolekce kvalifikačních prací nebo kolekce výzkumných prací.

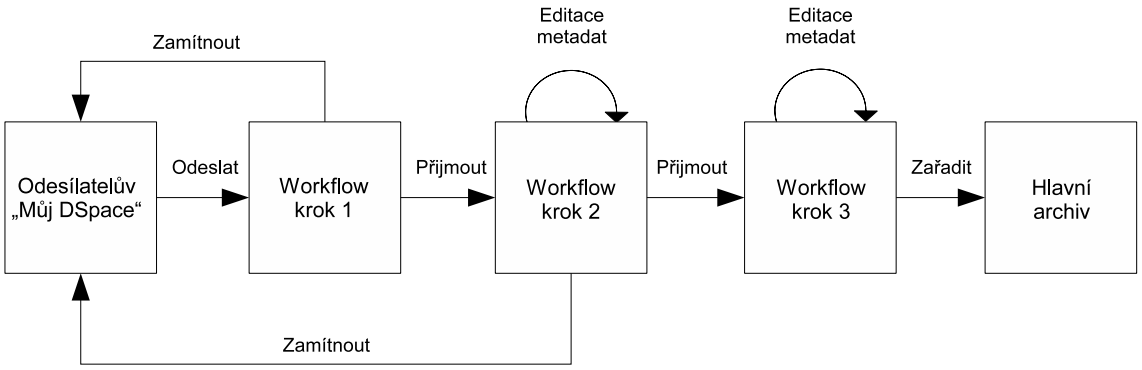
Každou položku vlastní jediná kolekce, ale položka se může objevit i v jiných kolekcích (nástrojem pro mapování položek). Každá položka je dále rozdělena do svazků souborů (bundles of bitstreams), kde každý svazek obsahuje jeden a více souborů. Svazkem se myslí například HTML dokument spolu s obrázky a styly nebo náhledy k obrázkům. Každý soubor je asociován s jedním formátem souboru (bitstream format).

3.4.3 Manažer datového skladu

Manažer datového skladu poskytuje univerzální rozhraní pro ukládání souborů. V současnosti nabízí dvě možnosti ukládání dat. První možností je ukládání na lokální soubo-



Obrázek 3: Datový model DSpace



Obrázek 4: Workflow proces

rový systém, kde úložiště může být rozděleno mezi několik fyzických disků nebo může být vše na jednom disku. Druhou možností je použití systému *storage resource broker* (SRB).

Storage resource broker je robustní systém poskytující téměř neomezenou diskovou kapacitu a mnoho možností pro zálohování dat. V tomto případě jsou data uložena buď na lokálním souborovém systému, nebo na vzdáleném serveru. Storage resource broker je podrobně popsán v [23].

3.4.4 Workflow proces – řízený koloběh dokumentů

Řízený koloběh dokumentů (workflow proces) je proces, který zpřehledňuje kontrolu, schvalování a přeposílání dokumentů. V DSpace tento proces zajišťuje kontrolu nad zadáváním položek do archívu, jejich schvalováním nebo kontrolu studentských prací jejich vedoucími. Workflow proces se skládá ze tří kroků, které mohou nebo nemusí být nastaveny. Každému kroku je přidělena skupina lidí, kteří mají na starost kontrolu a schvalování. Workflow proces je znázorněn na obrázku 4 (obrázek převzat z [2]).

Po odeslání registrovaným uživatelem se dokument přesune do tzv. zásobníku úloh pro první workflow krok. Tam se ho ujme některý z pověřených uživatelů, zkontroluje dokument a buď ho vrátí s příslušným vysvětlením autorovi nebo jej pošle do dalšího kroku. Pověřený uživatel pro druhý workflow krok pak může editovat metadata položky, případně položku pošle do dalšího kroku nebo ji vrátí zpět autorovi. Nemůže ovšem editovat samotný soubor s daty. Ve třetím kroku již jen pověřený uživatel zedituje metadata a pošle položku do hlavního archívu. Před přidáním položky do hlavního archívu se jí přiřadí trvalý identifikátor (handle), vyplní se datum, kdy se stala položka dostupná přes rozhraní DSpace, případně datum vydání. Dále se na položku aplikují výchozí politiky podle nastavení kolekce a položka se přidá do indexového souboru pro procházení a vyhledávání.

Pokud pro některý z kroků workflow procesu nemá nastavenu skupinu pověřených uživatelů, je jednoduše přeskočen a položka je automaticky poslána do dalšího kroku. Celý proces může přerušit administrátor DSpace přes webové rozhraní.

3.4.5 Dohlížení nad pracemi

DSpace nabízí vedoucím kvalifikačních prací možnost dohlížet na studenty a jejich práce. Přiřazování prací k dohlížení provádí administrátor. Každé práci je přiřazena skupina lidí, kteří na práci mohou dohlížet. Každé skupině je přiřazena výchozí politika, takže dohlížející může editovat zadání nebo jej jen sledovat a reagovat na změny prostřednictvím emailu. Tento mechanismus může být využit i pro spolupráci více lidí na jedné položce.

Výchozí politiky mohou být nastaveny na *editor*, *pozorovatel* nebo nejsou nastaveny vůbec a předpokládá se specifické nastavení politik administrátorem. *Pozorovatel* může prohlížet metadata i soubory s obsahem práce, ale nemá možnost v nich nic upravovat. *Editor* může pracovat s položkou, jako by byl jejím autorem.

3.4.6 Autorizační politiky

Autorizační politiky jsou v DSpace řešeny tak, že každé akci je přiřazen seznam lidí, kteří mohou akci provést. Takové seznamy jsou v DSpace nazývány *skupiny*. V systému jsou dvě speciální skupiny uživatelů. *Anonymní* skupina obsahuje všechny uživatele v systému a skupina *administrátoři* obsahuje správce systému, kterým je dovoleno dělat vše. Autorizační politiky lze přiřazovat komunitám, kolekcím, položkám, svazkům a souborům.
Komunitě můžeme přidělit oprávnění pro tyto akce:

- ADD – dovoluje přidat kolekce nebo subkomunity,
- REMOVE – dovoluje smazat kolekce nebo subkomunity.

Jednotlivým kolekcím můžeme přiřadit následující oprávnění:

- ADD – přidání položek do kolekce (může odesílat prostřednictvím formuláře),
- REMOVE – odstranění nebo změna kolekce,
- DEFAULT_ITEM_READ – všechny nově vložené položky budou mít tuto skupinu jako výchozí pro akci READ,
- DEFAULT_BITSTREAM_READ – všechny soubory vložené do kolekce budou mít tuto skupinu jako výchozí pro akci READ,
- COLLECTION_ADMIN – správcové kolekce, mohou editovat nebo mazat položky v kolekci nebo mapovat cizí položky do této kolekce.

Položky mohou mít oprávnění pro:

- ADD – přidávání svazků souborů k položce,
- REMOVE – odebírání svazků souborů z položky,
- READ – čtení položky (metadata jsou čitelná stále),
- WRITE – změna položky.

Svazkům můžeme přiřadit oprávnění pro:

- ADD – přidávání souborů do svazku,
- REMOVE – odstranění souborů ze svazku.

Jednotlivé soubory mohou mít jen oprávnění:

- READ – umožňuje číst soubory,
- WRITE – umožňuje zapisovat a modifikovat soubory.

3.4.7 Uchovávání čitelnosti formátu

DSpace rozděluje formáty souborů podle tří typů úrovně podpory – *podporovaný*, *známý* a *nepodporovaný*. Podle tohoto rozdělení pak se soubory pracuje. Hostitelská organizace může podle těchto úrovní zaručit převod mezi různými formáty.

Soubor typu *nepodporovaný* DSpace nerozpoznal a uchová jej v takové podobě, v jaké ho přijal. Ve stejné podobě jej poskytuje i uživatelům.

Pokud DSpace formát souboru rozpozná, může s ním zacházet dvěma způsoby. Jestliže soubor zařadí mezi typ *známý*, znamená to, že DSpace tento formát rozpoznal, uchová a poskytne jej ve stejné podobě, ale provozovatel DSpace nezaručuje převod do novějších formátů podobného typu. Mezi takové formáty patří například různé uzavřené formáty nebo formáty, které organizace nemůže nijak zpracovat (například chybí potřebné softwarové vybavení).

Jestliže provozovatel DSpace označí některý formát jako *podporovaný*, zaručuje, že formát tohoto typu bude čitelný i v budoucnosti. Toho může dosáhnout například převodem souborů do jiného formátu nebo emulací softwaru na zpracování. Do takové kategorie může provozovatel DSpace zařadit například obrázek ve formátu tiff, který je specifikovaný standardem ISO/IEC 10918-1 a existují pro něj konverzní metody.

3.4.8 Vyhledávání

DSpace nabízí uživatelům několik možností, jak najít požadovaný dokument. První možností je přístup prostřednictvím externího odkazu přímo na určitou položku, například pomocí identifikátoru handle.

Jednou z dalších možností, jak se dostat k objektu uloženém v DSpace je prohledávání popisných metadat. Jelikož jsou na vyhledávání kladeny vysoké nároky, vývojáři DSpace se rozhodli použít volně dostupný vyhledávací engine Apache Lucene [24]. Lucene poskytuje možnosti vyhledávání, indexování, přeskakování jazykových členů při

indexování a jednoduché přidávání záznamů do indexu bez potřeby přeindexovávat celý obsah. DSpace nabízí možnost jednoduchého vyhledávání, kdy se prohledávají všechna indexovaná metadata nebo rozšířené vyhledávání, kde lze určit i metadata, která se mají prohledávat. Rozšířené vyhledávání také nabízí možnost kombinovat vyhledávací výrazy pomocí logických operátorů OR, AND a NOT.

Neméně důležitou možností při hledání záznamů je procházení. DSpace dovoluje procházet názvy záznamů, jména jejich autorů a záznamy seřazené podle data vložení do DSpace. Je možno procházet celý obsah DSpace, jen záznamy patřící určité komunitě nebo záznamy patřící do určené kolekce. Při procházení je opět využit vyhledávací engine Lucene.

3.4.9 Licence Creative Commons

DSpace dává autorům možnost zvolit si alternativní licenci pro distribuci a zacházení s materiálem, který do úložiště vkládá. K tomuto je využito množství licencí, které poskytuje organizace Creative Commons [25]. Použití této licence je volitelné a krok zvolení alternativní licence může být přeskočen. Pokud uživatel využije možnost zvolit si svou licenci, je text této licence spolu s metadaty uložen jako soubor ve formátu RDF k ostatním souborům s obsahem.

3.4.10 Identifikace záznamů

Jak již bylo zmíněno dříve, DSpace používá k identifikaci digitálních objektů identifikátory typu CNRI handles [19]. CNRI handle system se stará o přidělování, rozlišování a manipulaci s identifikátory. Identifikátory jsou přiřazeny trvale, takže se znovu nepoužijí ani po odstranění záznamu z archívu. Pro reprezentaci identifikátorů je použit protokol HTTP a každé instalaci DSpace musí být přiřazen globálně jednoznačný prefix od hlavní autority.

V současné verzi DSpace jsou identifikátory přiřazovány komunitám, kolekcím a položkám. Soubory a svazky souborů mají své identifikátory odvozené od handlu. Handle může být zapsán dvěma způsoby:

1. hdl:4321.123/4567
2. http://hdl.handle.net/4321.123/4567

K rozpoznání identifikátoru zapsaného první možností je potřeba mít k dispozici nějaký rozlišovací software (například plugin pro webový prohlížeč nebo webovou aplikaci). Pokud zadáme do webového prohlížeče identifikátor zapsaný druhým způsobem, postará se o rozlišení server hdl.handle.net a přesměruje nás přímo na konkrétní instalaci DSpace. DSpace používá druhou formu zápisu. Handle server obsažený v DSpace se stará pouze o lokální část „4567“ a je pouze na něm, aby rozpoznal, zda jde o kolekci, komunitu nebo položku.

Každý soubor má v databázi přiřazen sekvenční ID, které je použito pro identifikaci souboru. Identifikátor pro soubor *text.pdf* odvozený od předešlého handlu by vypadal

například takto: „https://dspace.vsb.cz/bitstream/4321.123/4567/3/text.pdf“, kde „3“ je právě sekvenční ID.

3.4.11 Další

DSpace nabízí spoustu dalších možností a funkcí a jejich popis by tady byl zbytečný. Ve zkratce bych zmínil ještě několik důležitějších funkcí, které DSpace nabízí.

Důležitou funkcí pro spravování digitálního úložiště je export a import obsahu kolekcí. Export a import se používá k migraci systému DSpace na jiný server, ale také je to jedna z možností zálohování nebo obnovy dat. Jako výstupní formát exportu byl zvolen metadatový standard Dublin Core zapsaný v souboru XML. DSpace v současné verzi dokáže exportovat obsahy jednotlivých kolekcí a ovládání nástroje pro import a export je prováděno přes příkazový řádek. DSpace také dokáže exportovat data v metadatovém standardu METS [14].

Další užitečnou funkcí je tvorba statistik. DSpace rozděluje statistiky na měsíční souhrny, které mohou být veřejně přístupné nebo je přístup povolen pouze administrátorům. Statistiky obsahují tyto údaje:

- návštěvnost jednotlivých komunit, kolekcí a položek,
- počty přihlášení a odhlášení uživatelů,
- souhrn obsahu celého archívu (jaké typy záznamů jsou nejčastěji ukládány a další),
- nejčastěji vyhledávané slova a slovní spojení,
- setříděný seznam provedených akcí a některé další podrobné údaje.

DSpace dovoluje uživatelům registrovat se do kolekcí, u kterých mají zájem sledovat nové příspěvky. Těmto uživatelům je každý den zaslán email se seznamem nových příspěvků v kolekci. Pokud za předchozí den žádné příspěvky do kolekce nepřibyly, email se neposílá.

Pro zpracovávání obsahu archívu a také pro vytváření nového obsahu jsou v DSpace použity tzv. MediaFiltery. K základním filtrům patří extrakce textu uložených objektů pro fulltextové vyhledávání a tvorba náhledů k obrázkům ve formátu JPEG, GIF a PNG. Současné MediaFiltery vytvářejí z PDF souborů textové, sloužící k prohledávání obsahu. Tyto soubory ukládá s názvem *nazev.pdf.txt*. DSpace nabízí možnost doprogramovat si vlastní filtry pro jiné typy souborů.

Díky implementaci rozhraní a dobrého popisu těchto rozhraní je možné modifikovat a přidávat nové funkce implementací vlastních nástrojů, které využívají obecné rozhraní DSpace. O těchto úpravách se píše v kapitole 4.

3.5 Nasazení obecného digitálního repozitáře

Při budování digitálního repozitáře je vhodné postupovat podle určitých kroků. Tato kapitola popisuje jednotlivé kroky postupu. Každý digitální repozitář je svým způsobem

jiný a jsou na něj kladeny jiné požadavky (různé druhy digitálního obsahu, množství uchovávaných dat, modifikovatelnost a další).

Dříve, než začneme vybírat vhodný systém digitálního úložiště, je vhodné si rozmyslet jaké druhy digitálních objektů potřebujeme uchovávat, protože ne všechny systémy jsou vhodné pro uchovávání jakýkoliv typů dat a každý nabízí jiné služby. V našem případě potřebujeme uchovávat především naskenované články ze sborníků vědeckých prací, vysokoškolské kvalifikační práce a digitalizované staré tisky.

V další fázi bychom měli vybrat vhodný systém pro vybudování digitálního repozitáře. Měli bychom brát ohled na to, aby systém dokázal uchovat typy objektů, které požadujeme, a aby splňoval všechny požadavky, které jsme si stanovili v předchozím bodě. Měl by také umožňovat rozšíření typů ukládaných objektů a export obsahu v případě, že se rozhodneme přenést obsah úložiště do jiného systému. Zároveň bychom měli stanovit tým lidí, kteří budou digitální úložiště budovat, nastavovat a testovat. Měl bý to být tým složený z odborníků na informační systémy a knihovnické procesy. My jsme vybírali mezi systémy Eprints a DSpace, z nichž lépe vyhovuje DSpace.

Nyní může stanovený tým začít s instalací a úpravami repozitáře. Podle zvoleného systému, který může vyhovovat více či méně, provádíme patřičné úpravy. Protože oblast digitálních repozitářů se neustále vyvíjí, měli bychom při úpravách systému brát ohled na budoucí přechod na novější verze použitého systému a nemuseli tak novější verzi opět celou modifikovat.

Pokud je systém nainstalován a splňuje všechny požadované funkce, měli bychom stanovit, kteří uživatelé se budou o systém starat, kdo bude hlavním administrátorem a přidělit další role v systému. Těmto uživatelům by následně měla být přiřazena příslušná oprávnění. Dále by mělo následovat rozdělení obsahu repozitáře na jednotlivé skupiny lidí, kteří budou společně vytvářet nějakou kolekci dokumentů. Po vytvoření počátečních kolekcí bychom měli nastavit přístup k těmto kolekcím a určit, kdo bude moci do dané kolekce přispívat a kdo si ji jen prohlížet. Také bychom měli rozhodnout, jaké formáty dokumentů bude úložiště přijímat.

Současně s počátečním nastavováním úložiště musíme začít proškolovat běžné uživatele, kteří budou s úložištěm také pracovat. Případně můžeme vydat tištěný nebo elektronický návod k použití, který by měl obsahovat informace jak přispívat do úložiště, jak vyhledávat a přistupovat k obsahu a další.

V konečné fázi můžeme nasadit úložiště do ostrého provozu a začít digitální úložiště používat v praxi. Počáteční testování provozu by mělo odhalit všechny nedostatky, které se při testovacím provozu neprojevily. Měli bychom poskytnout uživatelům také podporu v případě potíží a můžeme zahájit nějaký druh propagační akce, který upozorní na nově vzniklé úložiště.

Celý tento postup bude v další kapitole podrobně popsán spolu s úpravami na námi zvoleném systému.

4 Řešení

V této kapitole bude popsán praktický postup při nasazování digitálního repozitáře v prostředí Ústřední knihovny Vysoké školy báňské – Technické univerzity Ostrava. Jak již bylo zmíněno dříve, my jsme se rozhodli jako systém digitálního úložiště použít DSpace. Budou zmíněny použité technologie a popsány jednotlivé úpravy, které bylo třeba provést v originálních zdrojových kódech a stránkách JSP. V další části následuje popis instalace DSpace na testovací server, lokalizace DSpace do češtiny, úpravy přihlašování prostřednictvím LDAP, úpravy vkládacích formulářů, nastavení přístupových práv a další úpravy. Dále je také popsán převod dat ze sytému T-Series do DSpace, možnosti zálohování a výsledky testování.

4.1 Popis technologií

Při implementaci DSpace byla použita řada technologií a standardních formátů, které usnadňovaly vývojářům při programování práci. At'už se jedná o použitou architekturu nebo zabezpečení přístupu a přenosu dat, bude třeba se blíže seznámit s těmito technologiemi. Pro podrobnější seznámení s nimi by bylo třeba popsat mnoho stránek, takže každá technologie bude stručně popsána alespoň tak, aby bylo možné správně pochopit následující úpravy.

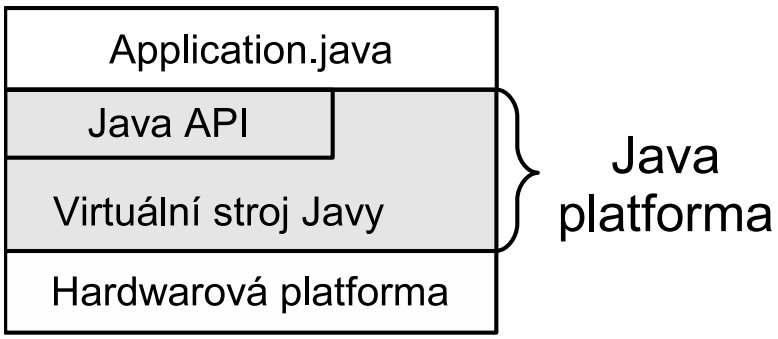
4.1.1 Java

Java [26] není jen programovací jazyk, ale je to celá softwarová platforma určená k vývoji aplikací. Největší uplatnění našla Java v oblasti internetových a multiplatformních aplikací. Program napsaný pro Java platformu můžeme spustit na řadě hardwarových platformách. Java platforma se skládá z programovacího jazyka Java a virtuálního stroje Javy.

Programovací jazyk Java je bezpečný, robustní, přenosný, objektově orientovaný jazyk, který není závislý na hardwarové platformě. Umožňuje tvorbu vícevláknových aplikací, použití síťových protokolů, a mnoho dalšího. Zdrojové kódy mají soubor s příponou *.java*, které se následně kompilují do podoby *byte code*, který má příponu *.class*.

Java platforma je znázorněná na obrázku 5. Většina platforem je tvořena kombinací hardwaru a softwaru, kdežto java platforma je pouze softwarovou platformou. Tvoří ji virtuální stroj Javy a rozhraní Java API. Java API je aplikační programové rozhraní jazyka Java. Program napsaný v Javě se nejdříve zkompiluje překladačem Javy do podoby tzv. *byte code*, který umí interpretovat virtuální stroj Javy. *Byte code* se dá chápat jako instrukční soubor pro virtuální stroj Javy. Kompilace aplikace se provede jen jednou na jakémkoliv stroji, ale spustit ji můžeme všude tam, kde máme k dispozici virtuální stroj Javy. Virtuální stroj Javy je software, který je dostupný pro mnoho platforem (například MS Windows, Linux, Unix, Solaris, ale také mobilní telefony a další zařízení).

Existuje několik verzí Java platformy. Pro tvorbu běžných aplikací na počítači je určena verze SE (Standard Edition), která obsahuje veškeré potřebné vybavení pro naprogramování (základní a rozšířené datové typy, práce se sítí, práce s databází, a další). Pro



Obrázek 5: Java platforma

tvorbu pokročilejších internetových aplikací je určena verze EE (Enterprise Edition), která přidává podporu internetových protokolů, webových aplikací (Java Server Pages, Java Servlets) a dalších. V dnešní době se stává stále populárnější verze ME (Micro Edition) pro mobilní zařízení jako jsou mobilní telefon nebo PDA. My jsme při implementaci použili J2SE (Java 2 Standard Edition) verze 1.5.0.

4.1.2 Java Server Pages

Java Server Pages [27] je technologie pro tvorbu dynamických webových stránek. Nabízí možnost rychlé tvorby webových aplikací, které jsou serverově a platformně nezávislé. V kombinaci s servlety jsou Java Server Pages výhodným nástrojem pro tvorbu rozsáhlých interaktivních aplikací běžících na straně serveru. Java Server Pages se nejčastěji používají pro tvorbu HTML dokumentů, které jsou následně zobrazovány v klientově prohlížeči.

Servlety jsou aplikace běžící na straně serveru, které doplňují stránky JSP. Používají se na rychlé a optimalizované zpracování dotazů. JSP stránky jsou po vygenerování serverem převedeny na servlety dočasně uložené na serveru a slouží pro rychlejší zpracování požadavků. Opakem servletu jsou applety, které běží na straně klienta a využívají virtuální stroj Javy na straně klienta.

JSP a servlety se dají využít pro oddělení statických a dynamických stránek. Statické stránky definují vzhled aplikace, zatímco dynamické stránky se servlety zajištují funkční logiku systému.

Hlavními přednostmi těchto technologií jsou:

- Definují jazyk pro tvorbu JSP stránek, což jsou textové dokumenty, definující jak přijmout, zpracovat a odeslat požadavek od klienta,
- definují mechanismy pro rozšíření jazyka JSP, přidávání nových tagů,
- nabízejí koncept pro přistupování objektů na straně serveru.

4.1.3 Tomcat

Apache Tomcat [29] je webový server, který slouží pro uchovávání a zpracovávání aplikacích založených na Java Servlets a Java Server Pages. Může být použit jako samostatný server nebo fungovat ve spojení s webovým serverem Apache. Tomcat je v podstatě skladiště servletů, které se stará o spuštění, běh a ukončení servletů a vyřizování požadavků od klientů pomocí protokolů HTTP nebo HTTPS. Výhodou Apache Tomcat je jeho dostupnost zdarma.

V našem případě byl použit Apache Tomcat ve verzi 5.5.15.

4.1.4 PostgreSQL

PostgreSQL je rozšířený, volně dostupný RDBMS (Relational DataBase Management System). Nabízí alternativu k jiným volně dostupným relačním databázím jako jsou MySQL a Firebird nebo ke komerčním databázovým systémům typu Oracle, MS SQL nebo DB2 od IBM. Podporuje mnoho moderních operačních systémů jako jsou MS Windows, Linux, OS/2. Je šířen pod licencí BSD [22], která umožňuje jeho modifikaci a binární distribuci.

Z jazyka SQL dovoluje použití cizích klíčů, vnořených dotazů, spojení tabulek (JOIN), spouští (trigger), pohledy (view) a dalších. Umožňuje psaní uživatelsky definovaných funkcí pomocí vestavěného jazyka PL/pgSQL nebo klasických programovacích jazyků jako jsou C, C++, Java, Perl, Tcl a další.

Naše digitální úložiště běží nad PostgreSQL verze 7.4.7 s ovladačem JDBC (Java Database Conectivity) pro programovací jazyk Java.

4.1.5 HTTPS

HTTPS (Hyper Text Transfer Protocol Secure) je protokol pro komunikaci prostřednictvím internetu. HTTPS zajišťuje bezpečný přenos dat za pomoci šifrováního protokolu Secure Socket Layer (SSL) nebo Transport Layer Security (TLS). HTTPS přenáší data protokolem HTTP, ale nepřenáší je v čistě textové podobě, ale šifrovaně. Toto šifrování znemožňuje odposlech přenášených dat a jejich podvržení. Komunikace protokolu HTTPS standardně probíhá na portu 443 (protokol HTTP komunikuje na portu 80).

Ověření identity serveru probíhá pomocí certifikátu serveru, který musí být elektronicky podepsán některou z certifikačních autorit. Webové prohlížeče a operační systémy v sobě mají napevno implementovány podepsané certifikáty hlavních certifikačních autorit, jako je například VeriSign [32]. Protokol HTTPS se nastavuje na úrovni webového serveru, který vyřizuje požadavky webové aplikace.

Komunikaci protokolem HTTPS poznáme podle adresy, která začíná řetězcem *https://* a celá adresa má tvar

```
https://hostitel:port/
```

kde *hostitel* je název serveru, na který se chceme připojit a *port* je číslo portu, na kterém přijímá požadavky aplikace nebo server. Pokud neuvedeme číslo portu, komunikuje se na standardním portu 443.

4.1.6 LDAP

Lightweight Directory Access Protocol (LDAP) je protokol určený pro udržování adresářů a práci s informacemi o uživatelích jako jsou vyhledávání adres, emailů a dalších informací uchovávaných v databázi nebo adresářové struktuře. Je založen na doporučení X.500 vyvinutý spolkem International Consultative Commitee of Telephony and Telegraphy (ITU-T), který definuje přenášení elektronických zpráv po počítačové síti. Konkrétní open source implementací je například OpenLDAP [30].

Je to protokol typu klient/server, takže klient se připojí k LDAP serveru (implicitně na port 389) a zašle požadavek, server mu pak vrací odpověď. Data v LDAP adresáři jsou řazena do stromové struktury a většinou popisují nějakou reálnou osobu, věc, tiskárnu nebo počítač. Každému takovému objektu je přiřazeno několik atributů (země, organizace, oddělení, jméno, email a další), nad kterými protokol LDAP pracuje. Ukázka z adresářové struktury školy je zachycena na obrázku 7.

Protokol umožňuje vyhledávat, přidávat, modifikovat a mazat záznamy v adresářové struktuře. Další podstatnou funkcí adresářové služby LDAP je možnost autentizace klienta. Autentizace pomocí LDAP se využívá v mnoha webových službách v rámci organizace, takže uživatelé používají pro přihlašování ke všem aplikacím stejné přihlašovací jméno a heslo.

4.1.7 XML

XML (eXtensible Markup Language) je značkovací jazyk podobný HTML (HyperText Markup Language). Hlavním rozdílem mezi těmito jazyky je skutečnost, že HTML svými značkami určuje, jak se mají data zobrazit, kdežto XML zachycuje význam popisovaných dat. XML byl navržen konsorciem W3C [31]. XML je stejně jako HTML odvozen od jazyka SGML (Standard Generalized Markup Language), který definuje popis a definici dat takovým způsobem, že nepracuje s daty jako s textem, ale jako s objekty. SGML je však příliš komplexní a složitý, a proto byl jako podmnožina odvozen jazyk XML. Jazyk XML byl navržen především jako rozšiřitelný jazyk, který povoluje dodefinování vlastních značek.

Jazyk XML je určen především pro výměnu dat mezi aplikacemi a popisování dokumentů. Snaží se popsat sémantiku dat místo toho, jak budou data prezentována. To zajišťuje propojení XML souboru s nadefinovanými styly, které určují, jak se má daný element zobrazit. Mezi nejpoužívanější styl pro zobrazování XML dokumentů na obrazovce nebo tiskárně patří CSS (Cascade Styles Sheets), který každému elementu přiřadí vlastnosti pro zobrazení (velikost písma, odsazení, barvu písma a další). Tato technika umožňuje stejný dokument zobrazit různě na monitoru a například na tiskárně. Pomocí transformačních stylů lze dokumenty v XML převádět do jiných formátů, jako jsou PDF, XHTML, HTML, postscript, a jiné.

Popsáním významu zachycených dat se naskýtá možnost efektivnějšího vyhledávání informací, kdy vyhledávače mohou prohledávat jen specifickou část obsahu. Například můžeme vyhledávat pouze text, který je obsažen v nějakém nadpisu, poznámce a jinde.

Na následující ukázce ze souboru XML je příklad zápisu článku. U článku je uveden autor, název a vlastní text. Každý XML soubor začíná deklarací XML. Ta musí obsahovat verzi XML a znakovou sadu, ve které je dokument napsán. Většinou se pro psaní XML používá znaková sada UTF-8, ale jsou povoleny i jiné. Následuje posloupnost značek (tagů) podle definice z DTD souboru. Značky jsou uzavřeny mezi znaky „<“ a „>“. Jelikož jsou tyto znaky použity pro označení značek, pro zapsání těchto znaků musíme použít předdefinovaných entit „<“ a „>“. Každá značka může obsahovat atributy a hodnoty nastavené těmto atributům. V příkladu je to například atribut *typ* a hodnota *vzdělávací* ve značce <*clanek*>.

```
<?xml version=”1.0” encoding=”UTF-8”?>
<clanek typ=”vzdělávací”>
   <autor>
      <jmeno>Lukáš</jmeno>
      <prijmeni>Jandera</prijmeni>
      <kontakt>lukas.jandera@tisk.cz</kontakt>
   </autor>
   <nazev>Jak získat státní občanství</nazev>
   <text>
      <nadpis>Úvod</nadpis>
      <odstavec>
         Tématem článku bude postup při získávání státního...
      </odstavec>
      ...
   </text>
</clanek>
```

Aby bylo možné zkontrolovat, jestli je daný dokument vytvořen správně, je třeba někde určit, jaké značky a v jakém pořadí se mohou v dokumentu vyskytovat. Popis takové struktury dokumentu je obsažen v DTD (Document Type Definition) souboru. DTD soubor popisuje schéma pro určitý typ XML dokumentu. Obsahuje seznam elementů a pro každý element také seznam možných vnořených elementů. Také určuje, zda-li je element povinný nebo nikoliv. Pro různé typy dokumentů byly vytvořeny různá DTD schémata, která můžeme použít nebo si můžeme vytvořit DTD s vlastním popisem XML souboru podle potřeby.

Aby byl dokument XML správně vytvořený, musí splňovat všechny následující pravidla:

- celý dokument musí být obsažen v jednom kořenovém elementu (root element),
- všechny elementy jsou párové, to znamená, že každý startovací element <značka> musí mít i svůj koncový element </značka> nebo pokud se jedná o prázdný element, může být ukončen značkou prázdného elementu <značka />,
- hodnoty atributů musí být uzavřeny v uvozovkách nebo v apostrofech, ikdyž jde jen o číselnou hodnotu,

- elementy se mohou do sebe zanořovat podle definice DTD, ale nesmí se křížit (<a>něco je špatný zápis),
- rozlišují se malá a velká písmena, takže startovací i ukončovací značka musí být zapsána stejně.

Oblasti použití XML jsou široké. XML se používá jako formát zápisu metadatových standardů, jako jsou METS, RDF nebo Dublin Core. Slouží také jako formát pro zápis a tvorbu publikací a spolu s předdefinovanými DTD je znám jako DocBook, který je podrobně popsán v kapitole 5.1. Je používán také pro popis multimédií a grafiky (dvourozměrná vektorová grafika SVG – Scalable Vector Graphics). Široké uplatnění našel XML formát také v komunikaci přes internet, například Instant Messenger Jabber je založen na zprávách ve formátu XML. Používá jej také mnoho aplikací pro výměnu dat mezi jednotlivými instalacemi.

4.2 Instalace

Instalace DSpace proběhla bez větších problémů, bylo ovšem potřeba před instalací připravit operační systém a nainstalovat potřebné softwarové vybavení. Pro instalaci je třeba zvolit dostatečně výkonný server, který by měl disponovat velkou diskovou kapacitou, velkou a rychlou pamětí RAM a rychlým procesorem. Vývojáři DSpace je doporučován server s minimálně 2 GB operační paměti, dvoujádrovým procesorem Xeon 2,4 GHz a SCSI disky s velkou kapacitou. Pro náš testovací DSpace byl vybrán počítač s procesorem Intel Pentium 4 běžícím na 2 GHz a 512 MB operační paměti.

Při instalaci jsem spolupracoval s ing. Stanislavem Ulmanem, který zajišťoval správu systému a poskytl již předinstalovaný počítač. Na tomto testovacím počítači byl nainstalován operační systém Debian GNU/Linux verze 3.1, Java 2 Runtime Environment verze 1.5 a databázový relační systém PostgreSQL verze 7.4.7. Následně byl doinstalován ještě program *ant*, který je potřeba při kompilování a sestavování projektu DSpace. Po vytvoření databáze *dspace* a nastavení příslušných oprávnění jsem mohl přistoupit k samotné instalaci.

Ze serveru sourceforge.net jsem stáhnul zdrojové kódy DSpace verze 1.3.2 a rozbalil je do svého adresáře. V konfiguračním souboru *dspace.cfg* jsem nastavil údaje potřebné pro instalaci, jako jsou adresář, kam se bude instalovat, jméno a heslo databáze, adresu poštovního serveru a adresu, na které DSpace poběží. Do adresáře se zdrojovými kódy jsem nakopíroval ovladač JDBC pro PostgreSQL příslušné verze. Dále jsem pomocí programu *ant* spustil kompilaci a instalaci zdrojových kódů DSpace. Po úspěšné kompilaci jsem přesunul vytvořené zabalené archívy aplikace do správného adresáře serveru a tomcat restartoval. Poslední krok základní instalace je vytvoření administrátora DSpace pomocí příkazu *create-administrator*. Ostatní administrační úkony lze provádět pomocí webového rozhraní DSpace.

Po nainstalování a spuštění základní instalace bylo potřeba zajistit pravidelné spouštění některých akcí. Toho bylo dosaženo pomocí linuxového démona *cron*, který ve stanovený čas spouští zadané příkazy. Cron spouští skripty pro rozesílání pravidelných emailů, čištění databáze od položek označených jako smazané, skript *media-filter* pro extrahování

textu z vložených dokumentů a vytváření náhledů obrázků. Každý den také provádí generování statistik přístupů a dalších údajů, jak již bylo popsáno v kapitole 3.4.11. Jednou týdně se spouští skript pro fyzické mazání souborů, které byly označené za smazané.

Po takto nainstalovaném DSpace se objevily některé problémy při vyhledávání a procházení českých názvů. Problémy se projevovaly dvojím způsobem.Při vyhledávání se některé znaky v hledaném řetězci (většinou se jednalo o české znaky s diakritikou) po odeslání požadavku změnily na nezobrazitelné znaky a výsledky hledání tím byly ovlivněny. Příčinou byly špatně nastavené parametry při spouštění serveru Tomcat, konkrétně použitá znaková sada pro kódování textu. Po nastavení kódování na znakovou sadu UTF-8 problém s vyhledáváním zmizel. Dalším problémem bylo třídění českých názvů. Názvy se netřídily podle pravidel pro český jazyk, takže názvy začínající písmenem s diakritikou se zařazovaly na začátek seznamu a písmeno „Ch“ bylo vřazeno mezi názvy začínající na „C“. Problém ovšem nebyl v DSpace, ale ve špatně nastavené databázi PostgreSQL. Protože PostgreSQL byl instalován a inicializován již při instalaci systému, který byl nastaven pro anglické prostředí, třídění dat se provádělo podle těchto počátečních nastavení. Pro opravení bylo nutné data z databáze zálohovat do souboru pomocí příkazu *pg_dump*, smazat obsah celé databáze a poté databázi znovu inicializovat se správně nastavenými parametry. Příkaz pro inicializaci databáze s nastavením českého třídění vypadá následovně:

```
initdb -D [postgres]/data --locale=cs_CZ.UTF-8
```

kde [postgres] je adresář, ve kterém mají být data uložena. Po této opravě již DSpace fungoval korektně a mohl jsem začít s modifikacemi pro potřeby knihovny.

Po instalaci byly vytvořeny základní komunity a kolekce. Komunity jsou rozděleny podle vysokoškolských fakult a pracovišť. V současné době jsou v DSpace kolekce dvojího typu, vysokoškolských kvalifikačních prací a naskenovaných článků ze sborníků vědeckých prací. Pod komunitou „Ústřední knihovna“ je kolekce s importem vysokoškolských kvalifikačních prací ze systému T-Series. Připravuje se také kolekce pro digitalizované staré tisky. Nyní (květen 2006) je DSpace instalován na testovacím serveru v knihovně na adrese *http://pcnk233c.vsb.cz:8080/dspace*.

Více podrobností je o instalaci a nastavení DSpace popsáno v [2] nebo v administrátorské dokumentaci, která je v elektronické podobě přiložena na CD k této práci.

4.3 Lokalizace do češtiny

Důležitou úpravou, která byla požadována, je lokalizace systému DSpace do českého jazyka. Vývojáři již začali s přípravou na internacionalizaci prostředí, takže tato úprava nebyla tak složitá. V současné verzi DSpace je plně internacionalizované webové prostředí systému, proto stačilo pouze přeložit soubor s texty a popisy.

Při internacionalizaci jsou použity prostředky platformy Javy, které toto umožňují docela snadno. Pro texty v JSP stránkách byla použita knihovna Java Standard Tag Library a pro výběr správného souboru s texty byla použita třída *ResourceBundle* z Java API. Soubory s texty jsou pojmenovány podle vzoru „Messages_cs.properties“, kde cs je kód jazyka podle normy ISO 639. Pokud nějaký text není nalezen v lokalizovaném souboru,

Obrázek 6: Upravený index písmen pro procházení

použije se anglický text z výchozího souboru *Messages.properties*. Přepínání mezi jazyky je prozatím vyřešeno na základě detekce nastavení webového prohlížeče klienta. Pokud jazyk nastavený v prohlížeči nemá v DSpace odpovídající soubor s texty, DSpace se zobrazí s anglickými texty z výchozího souboru.

Výchozí soubor s texty *Messages.properties* je na následující ukázce. Na začátku řádku je vždy uveden klíč, podle kterého se rozpoznávají jednotlivé texty a za znakem „=“ je uveden text, který se zobrazí na příslušné stránce. Název klíče je odvozen od plné cesty k JSP stránce a řetězce identifikujícího text v rámci jedné stránky.

```
jsp.layout.navbar-default.about = About DSpace
jsp.layout.navbar-default.advanced = Advanced Search
jsp.layout.navbar-default.authors = Authors
jsp.layout.navbar-default.browse = Browse
```

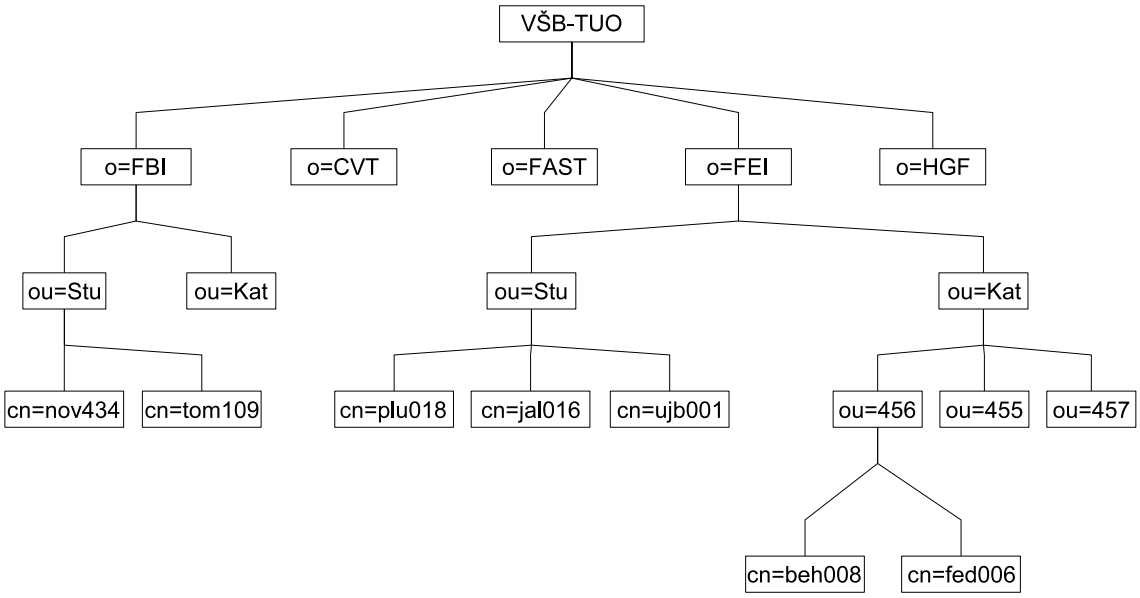
Soubor s texty musí být v kódování ascii, takže české znaky musí být zapsány jako sekvence „\uXXXX“, kde XXXX je hexadecimální číslo určující znak v ascii tabulce. Soubor jsem přeložil v kódování iso-8859-2 a následně jej překonvertoval do ascii pomocí příkazu *native2ascii*. Po přeložení textů z předešlé ukázky vypadají tyto texty následovně:

```
jsp.layout.navbar-default.about = O DSpace
jsp.layout.navbar-default.advanced = Rozšířené hledání
jsp.layout.navbar-default.authors = Autoři
jsp.layout.navbar-default.browse = Procházet
```

Soubory s nápovědou nejsou zatím internacionalizovány, takže prozatím zůstaly v angličtině. Přeloženy byly šablony emailů, které sice také nejsou přizpůsobeny vícejazyčné verzi, ale DSpace rozesílá tyto emaily uživatelům po vložení záznamu do repozitáře, pokud se uživatel přihlásí k odběru emailů s novými záznamy vloženými do kolekce nebo v jiných případech. Tyto emaily se tedy posílají pouze v české verzi, nezávisle na tom, jak je nastaven webový prohlížeč.

Přeloženy byly také formuláře pro vkládání záznamů do DSpace, které sice nejsou součástí lokalizačního souboru, ale předpokládá se nejčastější nastavení pro český jazyk. Souběžně s lokalizací formulářů byly provedeny úpravy formulářů tak, aby splňovaly požadavky na vkládání požadovaných dokumentů. Tyto úpravy budou podrobněji popsány dále.

V rámci lokalizace DSpace do českého jazyka byla upravena JSP stránka pro procházení názvů a autorů, kde byl upraven seznam počátečních písmen, do kterého byla přidána písmena „Ch“, „Č“, „Ř“, „Š“ a „Ž“. Tento seznam je na obrázku 6.



Obrázek 7: Adresářová struktura LDAP

4.4 LDAP

Dalším požadavk na digitální úložiště byla možnost přihlásit se prostřednictvím školního LDAP uživatelského jména a hesla. Uživatelé se nemusí registrovat, pokud chtějí vložit do DSpace svou práci. Navíc se nabízí možnost odlišit uživatele přihlášené pomocí LDAP hesla od uživatelů přihlášených přes svůj registrovaný účet v DSpace.

V základní instalaci nabízí DSpace možnost přihlášování přes LDAP. Vyskytl se ovšem problém při nastavení údajů pro ověřování uživatelů. Byly nastaveny údaje jako adresa LDAP serveru, názvy polí v adresářové struktuře a kontext pro ověřování uživatelských hesel. DSpace však předpokládá, že uživatelé jsou v adresářové struktuře LDAP v jednom podstromě. V nastavení je nastaven kontext, ve kterém DSpace uživatele hledá a pokud je najde, ověří heslo. Na obrázku 7 je ukázka adresářového stromu školního LDAP serveru (ldap://ldap.vsb.cz:389). Každá fakulta má svůj podstrom jako samostatná organizace, takže například *o=FEI* znamená „organization FEI“. Fakulty jsou dále rozděleny na studenty, katedry a další. Pod katedrami je vždy seznam kateder podle čísla katedry a v nich jednotliví zaměstnanci kateder. Pod studenty (*ou=Stu*) je seznam všech studentů fakulty. Pokud byl kontext nastaven například na „o=FEI,ou=Stu“, mohli se přihlašovat pouze studenti fakulty elektrotechniky a informatiky.

Problém jsem vyřešil tak, že jsem implementoval vlastní servlet pro přihlašování – *VSB_LDAPServlet.java*, který doplňoval funkce původního servletu. Servlet nejdříve prohledá celý adresářový strom a snaží se najít uživatele s příslušným uživatelským jménem. Pokud takového uživatele najde, použije kontext z výsledků hledání pro ověření uživa-

telského hesla. Servlet jsem přidal k ostatním a v souboru *dspace-web.xml* s nastavením webové aplikace jsem jej zaměnil místo původního servletu *LDAPServlet.java*.

DSpace nabízí programové rozhraní pro vytvoření vlastních autentikačních pravidel. Rozhraní umožňuje implementovat tyto metody:

- *allowSetPassword(context, request, email)* – pro konkrétního uživatele vrací *true* nebo *false*, pokud si může uživatel sám nastavit heslo,
- *canSelfRegister(context, request, email)* – pro konkrétního uživatele vrací *true* nebo *false*, pokud se uživatel může zaregistrovat do DSpace a vytvořit si tak vlastní účet, přístupný přes svoji emailovou adresu,
- *getSpecialGroups(context, request)* – vrací seznam speciálních (dynamických) skupin, do kterých je uživatel zařazen. Skupiny se mohou tvořit na základě emailové adresy, IP adresy nebo dalších údajů,
- *initEperson(context, request, eperson)* – předvyplnění informací o uživateli, pokud se uživatel automaticky registruje například pomocí LDAP,
- *startAuthentication(context, request, response)* – spouští servlet pro autentikaci.

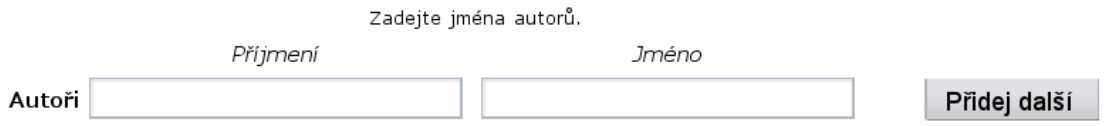
V našem případě byla implementována hlavně metoda *getSpecialGroups()*, která uživatele, kteří mají email končící „@vsb.cz“, přidává do skupiny VSB_users. Těmto uživatelům přihlášeným přes LDAP metoda *allowSetPassword()* nedovoluje změnu hesla. Ostatní metody rozhraní, které nebylo třeba modifikovat, byly převzaty z výchozí třídy *SimpleAuthenticator.java*.

4.5 Vkládací formuláře

Aby bylo možné vkládat do DSpace různé typy digitálních dokumentů, musely být vytvořeny odpovídající formuláře pro vkládání nových záznamů. DSpace má formuláře uloženy v jednom XML souboru, ve kterém je také uloženo, jaký formulář má mít konkrétní kolekce. Soubor se jmenuje *input-forms.xml* a je načítán vždy při spuštění DSpace. Jak již bylo zmíněno dříve, tento soubor byl lokalizován do českého jazyka a byly do něj přidány další formuláře. Na následující ukázce je zobrazena struktura souboru *input-forms.xml*.

```
<input-forms>
 <form-map>
  <name-map collection-handle="default" form-name="trad" />
  ...
 </form-map>

 <form-definitions>
  <form name="trad">
  ...
 </form-definitions>
```



Obrázek 8: Ukázka políčka formuláře

```
 <form-value-pairs>
  <value-pairs value-pairs-name="languages" dc-term="language">
  ...
 </form-value-pairs>
</input-forms>
```

Na začátku jsou mapovány formuláře na jednotlivé kolekce a je tady uvedeno také výchozí mapování, které přiřadí formulář *trad* všem kolekcím, které nejsou v tomto seznamu. Dále následují definice pojmenovaných formulářů, které v sobě obsajují jednotlivé formuláře, stránky formulářů a políčka. Nakonec jsou uvedeny seznamy hodnot pro prvky formuláře umožňující výběr z předdefinovaných hodnot.

Každý prvek formuláře má odpovídající element v definici formuláře *form*. Jeden prvek formuláře mohou definovat následující značky v XML souboru *input-forms.xml*:

- *dc-element* – povinný element určující Dublin Core element, do kterého se hodnota zapíše,
- *dc-qualifier* – určuje upřesňující Dublin Core kvalifikátor, do kterého se zapíše hodnota z prvku formuláře,
- *repeatable* – povoluje opakovatelnost políčka formuláře. U políčka formuláře se vygeneruje tlačítko pro přidání dalšího políčka stejného typu,
- *label* – povinný element pro název políčka,
- *input-type* – povinný element pro určení typu formulářového políčka. Může to být textové pole, dvojité pole pro jméno a příjmení, speciální trojité pole pro datum nebo combobox pro výběr z přednastavených hodnot,
- *hint* – povinný nápovědný text pro upřesnění obsahu, zobrazuje se nad formulářovým políčkem,
- *required* – nastavuje povinně vyplnitelné políčko. Pokud bude chtít uživatel přejít na další stranu vkládacího formuláře, zobrazí se mu upozornění o nevyplněném poli.

Na obrázku 8 je zobrazen jeden prvek formuláře, který se vygeneroval z následující části souboru *input-forms.xml*:

```
<field>
   <dc-element>contributor</dc-element>
   <dc-qualifier>author</dc-qualifier>
   <repeatable>true</repeatable>
   <label>Autoři</label>
   <input-type>name</input-type>
   <hint>Zadejte jména autorů.</hint>
   <required>Nevyplnili jste žádného autora!</required>
</field>
```

Vyplněný autor se uloží do Dublin Core elementu „contributor.author“ a pokud nebude vyplněn žádný autor, zobrazí se upozorňující text z elementu *required*.

Bylo vytvořeno několik typů formulářů pro vysokoškolské kvalifikační práce, články ze sborníků, staré tisky a několik obecných formulářů. Formulář pro vysokoškolské kvalifikační práce byl nakonec rozdělen do sedmi formulářů podle fakult a každému formuláři byl nastaven jiný předvolený seznam studijních oborů, programů a kateder.

4.6 Zabezpečení přístupu

Zabezpečení přístupu lze rozdělit na dvě hlediska. Z prvního hlediska je nutné zabezpečit komunity a kolekce v DSpace tak, aby do nich nemohli vkládat všichni uživatelé, ale jen ti, kterým je to povoleno. Dále je třeba zabezpečit samotný přístup k serveru po síti, aby neoprávnění uživatelé nemohli odposlechnout heslo při přihlašování jiného uživatele.

4.6.1 Přístupová práva

Základním požadavkem na přístupová práva bylo omezení přístupu k plným textům skenovaných článků ze sborníku vědeckých prací. Jak již bylo zmíněno v kapitole 4.4, uživatelé přihlášení pomocí svého školního LDAP jména a hesla se automaticky řadí do dynamické skupiny *VSB_users*. Oprávnění pro čtení souborů v těchto kolekcích bylo nastaveno skupině *VSB_users*. Ostatní uživatelé mohou číst pouze metadata těchto záznamů. Vkládání do této kolekce je povoleno jen pověřeným knihovníkům. Vkládání do kolekcí vysokoškolských kvalifikačních prací je prozatím povoleno pouze pověřeným knihovníkům, ale předpokládá se, že v budoucnu budou do DSpace vkládat své kvalifikační práce samotní studenti, tedy uživatelé ve skupině *VSB_users*.

Aby bylo možné kontrolovat správnost vkládaných prací a doplnění knihovnických údajů, bude nutné nastavit pro všechny kolekce kvalifikačních prací proces schvalování a proces finální editace metadat, kterou bude mít na starost pověřený knihovník.

4.6.2 Zabezpečení serveru

Zabezpečení serveru spočívá v šifrování dat přenášených po síti pomocí protokolu HTTPS (viz kapitolu 4.1.5). DSpace je navrhnut tak, aby bylo možné tento protokol použít. Pro šifrovaný přístup stačí nastavit kontejnér pro servlety Apache Tomcat. Nastavení serveru Apache Tomcat pro zprovoznění šifrovaného přenosu probíhá ve třech krocích:

1. je potřeba vygenerovat certifikát, který zaručí správnou identitu serveru. Certifikát se vygeneruje pomocí nástroje *openssl*,
2. certifikát se odešle certifikační autoritě k podpisu,
3. nastavit správně server Apache Tomcat. V souboru *server.xml* nastavíme sekci *Connector* pro zabezpečené připojení. Musíme také nastavit port, na kterém bude server přijímat požadavky přes SSL. Standardně je to port 8443.

Takto nastavený Tomcat je přístupný například přes adresu *https://pcnk233c.vsb.cz:8443/*. Pokud bychom chtěli použít Tomcat bez specifikování portu, museli bychom použít buď přesměrování pomocí *iptables* v linuxu nebo přesměrování požadavků na Tomcat přes webový server Apache.

Až bude DSpace nasazen v Ústřední knihovně do ostrého provozu na hlavní server, bude použito přesměrování přes webový server Apache a adresa může vypadat například takto: *https://dspace.vsb.cz/*.

4.7 Ostatní úpravy

Při testovacím provozu bylo odhaleno ještě mnoho nedostatků a návrhů na vylepšení, takže se DSpace ještě postupně upravoval za provozu. Byl upravován vzhled DSpace, velikosti písem některých textů, zobrazování výsledků procházení podle počátečního písmena a další. Těchto úprav bylo mnoho, proto tady zmíním jen ty podstatnější.

4.7.1 Úprava metadat

Aby bylo možné do DSpace vkládat také záznamy o vysokoškolských kvalifikačních pracích, bylo potřeba rozšířit standardní Dublin Core metadata o několik nových položek. Tato metadata jsou zapsána v XML souboru, který se při instalaci přenese do databáze. Po nainstalování lze metadatové registry editovat prostřednictvím administrátorského rozhraní DSpace. Tato nově přidaná metadata musela být zahrnuta také do vkládacích formulářů. V tabulce 2 je seznam přidaných metadat do Dublin Core registrů.

4.7.2 Vyhledávání

Protože byly dodefinovány vlastní prvky metadat, bylo třeba tyto metadata zahrnout do rozhraní pro rozšířené vyhledávání. Mimo těchto přidaných jsme potřebovali přidat i některé stávající prvky Dublin Core. Ve finální podobě byly do rozšířeného vyhledávání zařazeny tyto možnosti:

- *Autor* – vyhledává ve jménech autorů, vedoucích prací a oponentů,
- *Název* – prohledává jakékoliv slovo z názvů,
- *Klíčové slovo* – hledá v klíčových slovech,
- *Abstrakt* – prohledává český i cizojazyčný abstrakt dokumentů,

DC element	DC kvalifikátor	Význam
contributor	consultant	Konzultant práce
contributor	referee	Oponent práce
description	abstract-en	Cizojazyčný abstrakt práce
date	accepted	Datum obhájení práce
thesis	degree-name	Jméno přidělované hodnosti
thesis	degree-level	Typ studijního programu
thesis	degree-branch	Studijní obor
thesis	degree-program	Studijní program
thesis	degree-grantor	Instituce přidělující hodnost
description	department	Katedra
description	category	Kategorie práce
identifier	location	Lokace práce
identifier	signature	Signature práce

Tabulka 2: Přidané prvky metadat

- *Zdrojový dokument* – hledá citaci zdrojového dokumentu,
- *Identifikátor* – jakýkoliv identifikátor jako handle, signatura, přírůtkové číslo a podobně,
- *Kód jazyka* – vyhledá dokumenty napsané v zadaném jazyce,
- *Druh dokumentu* – omezení na druhy prací jako jsou diplomové, bakalářské, a další,
- *Datum* – jakýkoliv datum (vydání, zveřejnění v DSpace a jiné),
- *Studijní obor/program* – hledá v názvech studijních oborů a programů,
- *Instituce/katedra* – prohledává názvy kateder a institucí.

Pro zajištění těchto úprav bylo třeba upravit JSP stránku pro rozšířené vyhledávání a upravit nastavení Lucene search indexů pro vyhledávání.

4.7.3 Modifikace servletu pro vkládání

Při testování DSpace vznikl požadavek na přeskočení nahrání souboru s digitálním dokumentem. Toho může být využito například pro vložení metadat kvalifikačních prací nebo článků, které ještě nejsou digitalizovány. Pro dosažení takového chování musel být upraven servlet pro vkládání *SubmitServlet*, který zajišťuje kontrolu nahraného souboru. Dále bylo třeba upravit několik JSP stránek s vkládacím formulářem.

Při úpravě byla přidána do souboru *input-forms.xml* s definicemi formulářů možnost volby *fileupload="optional"*, která určuje, zda může být nahrání souboru přeskočeno a uloženy budou pouze metadata. Pokud tato volba v definici formuláře je, na stránce formuláře, kde se nahrává soubor, se objeví tlačítko pro přeskočení nahrání. Soubor pak může být doplněn knihovníkem nebo administrátorem.

4.8 Převod dat z T-Series

Hlavním úkolem práce bylo převedení stávajících metadat o vysokoškolských kvalifikačních pracích a naskenovanými články ze sborníku vědeckých prací ze systému T-Series. Systém T-Series podporuje několik formátů exportů, ale ani jeden nebyl přímo použitelný pro import pomocí nástrojů dodávaných s DSpace. DSpace dokáže importovat data zapsaná v XML souboru standardu Dublin Core. T-Series umí exportovat do značkovaného textového souboru, který se zdál být nejvhodnější pro import. Bylo ovšem potřeba vytvořit konverzní program, který metadata z textového souboru převede do XML souboru ve standardu Dublin Core.

Konverzní program byl vytvořen v jazyce Java, jako vstup přijímá exportovaný soubor z T-Series v kódování UTF-8 a výstupem je adresářová struktura s XML soubory připravenými pro import do DSpace. Program také vypisuje záznamy, u kterých nebyla vyplněna některá metadata a vytváří soubor se seznamem importovaných PDF souborů. Exportované soubory z T-Series jsou v kódování 852, takže byly do UTF-8 převedeny pomocí linuxové utility *iconv*. Aby bylo možné rozpoznat, který PDF soubor patří kterému metadatovému záznamu, byl použit XML soubor ze systému pro převod naskenovaných stránek článků do PDF, který implementoval jako svou diplomovou práci Jan Vitásek [4]. V tomto souboru (*books.xml*) je seznam článků s jejich metadaty a také s názvem PDF souboru s naskenovaným textem. Pro určení správného souboru bylo třeba porovnat čtyři hodnoty – název sborníku, číslo sborníku v roce, rok a referenční číslo.

Soubor se seznamem rozpoznaných PDF sloužil ke kontrole počtu importovaných souborů a dá se z něj pomocí několika linuxových utilit vytvořit také soubor se seznamem nerozpoznaných PDF souborů. Z celkového počtu asi 800 naskenovaných článků zůstalo nerozpoznáno asi 70 souborů, protože nebyly zapsány v rozpoznávacím souboru *books.xml*. Ze seznamu těchto 70 nerozpoznaných souborů byl vytvořen HTML dokument s odkazy na PDF soubory, aby mohli knihovníci zkontrolovat a případně ručně přidat soubory k metadatům v DSpace.

Výstupní adresář, do kterého konverzní program zapisuje převedené Dublin Core metadata a rozpoznané PDF soubory, musí mít přesně danou strukturu, aby bylo možné bezchybně importovat záznamy do DSpace. Každý záznam pro import je v samostatném podadresáři. Tento adresář obsahuje následující soubory:

- *dublin_core.xml* – soubor s metadaty Dublin Core,
- *contents* – obsahuje seznam souborů, které mají být při importu přiřazeny k záznamu. Na každém řádku je jedno jmého souboru,
- *text.pdf* – soubor s naskenovaným textem. Takových souborů může být přidáno několik, při importu budou do DSpace nahrány všechny.

4.8.1 Popis konverzního programu

V první fázi program pouze převáděl formát souboru z textového do XML. Jak se ale ukázalo později, bylo třeba provést s daty při převodu některé úpravy. Jednou z úprav

bylo vyfiltrování nežádoucích údajů v exportovaných datech. Pro zajištění jednoznačnosti se v systému T-Series k některým údajům přidávaly například signatury, názvy nebo rok narození autora. Všude se používala zpětná lomítka, která dávala slovům nebo znakům v T-Series speciální význam. Tyto nežádoucí údaje a lomítka bylo třeba správně vyfiltrovat, aby v nových datech byly jen správné údaje. Navíc se se změnou uživatelů vkládajících do T-Series záznamy měnily i zvyky a způsoby zápisu, takže se s tím muselo při převodu počítat.

Další úpravou bylo přidání ISSN a vydavatele u článků ze sborníku na základě roku vydání a názvu sborníku. U těchto článků se navíc hledal i naskenovaný text, jak už bylo zmíněno dříve.

Program je složen z těchto tří tříd:

- *Record.java* – třída implementující jeden záznam se všemi údaji, která dovede záznam zapsat do XML souboru a provádí úpravy a filtrace údajů,
- *Converter.java* – třída, která provádí konverzi dat tím, že čte vstupní soubor a rozpoznává v něm jednotlivé záznamy. Tyto záznamy pak vkládá do instance třídy *Record* a vytváří výstupní adresářovou strukturu. Pro soubor obsahující články načítá XML soubor s údaji o naskenovaných souborech a hledá v něm jméno souboru pro daný článek. Pokud soubor nalezne, zkopíruje jej do výstupního adresáře,
- *Conv.java* – spouštěcí třída, která tvoří uživatelské rozhraní pro práci s konverzním programem. Zajišťuje rozpoznávání a kontrolu vstupních parametrů a případný výpis nápovědy. V současné verzi poskytuje rozhraní pro spouštění z příkazové řádky, ale může být rozšířena o implementaci grafického rozhraní nebo může být nahrazena servletem, který může být přidán do administrátorského rozhraní DSpace a umožnit tak snadný převod dat ze systému T-Series i s importem.

Podrobnější programátorská dokumentace k programu je k dispozici na přiloženém CD. Dokumentace je generována ze zdrojových kódů technologií javadoc, jak je u jazyka Java zvykem.

4.8.2 Ukázka převedeného záznamu

Následuje názorná ukázka z převodu článků sborníku vědeckých prací. Pro ukázku jsem vybral jeden z asi 3500 článků. Každý záznam začíná nezobrazitelným znakem „Line Feed“, podle kterého rozpoznám jednotlivé záznamy v jinak dlouhém neoddělaném souboru. Na jednom řádku je vždy zapsán jeden údaj, který začíná značkou identifikující, o jaký údaj jde. Za čtyřznakovou značkou následuje dvojtečka a za ní samotný textový údaj. Z řádku se značkou *ADRN* je použito referenční číslo v hranatých závorkách pro nalezení naskenovaného souboru.

```
ATIT:Testing and statistical feedback
ARES:Petr Šaloun, Dana Šalounová, Anna Madryová
A/TI:Testing and statistical feedback
AOTI:Testování a statistická zpětná vazba
```

```
A/OT:Testování a statistická zpětná vazba
AAAX:Šaloun, Petr,\\\ 1962-@Testing and statistical feedback
AAUT:Šaloun, Petr,\\\ 1962-
AAAX:Šalounová, Dana,\\\ 1963-@Testing and statistical feedback
AAUT:Šalounová, Dana,\\\ 1963-
AAAX:Madryová, Anna,\\\ 1952-@Testing and statistical feedback
AAUT:Madryová, Anna,\\\ 1952-
ADRN:článek 27 [3]
ALAN:anglicky \\eng\
ASER:Sborník vědeckých prací Vysoké školy báňské - Technické
univerzity Ostrava.\\,\ Řada elektrotechnická \\a\
A/ST:Sborník vědeckých prací Vysoké školy báňské - Technické
univerzity Ostrava.\\,\ Řada elektrotechnická \\a\
ASNR:Roč. 5, č. 1
ASPG:s. 23-32 : il.
ASYY:1999
AKWD:Feedback
AKWD:Statistical
AKWD:Testing
AEDT:2001/03/22
AUDT:2003/10/02
AENT:OSD 002/HAU50
AUPD:OSD 002/HAU50
```

Na obrázku 9 je ukázka souboru *dublin_core.xml*, kde je již vidět výsledný XML soubor konverze. Lze v něm vidět vyfiltrování lomítek a dat narození autorů, přidaný údaj o vydavateli a ISSN a citace poskládaná z několika údajů původního souboru. Podle klíčových slov cizích jazyků byl rozpoznán jazyk názvu, který je důležitý pro správné řazení cizojazyčných názvů, u kterých se jako počáteční slovo neberou členy jazyka. Podle prvního písmena značky v předchozím souboru byl rozpoznán typ záznamu jako „článek“.

Na obrázku 10 je vidět, jak tento záznam vypadá po naimportování do DSpace. Ve výpisu záznamu se vypisují jen některé údaje, plný výpis se zobrazí až po kliknutí na „Zobraz celý záznam“. Z obrázku je také vidět, že při konverzi nebyl nalezen odpovídající naskenovaný PDF soubor a záznam je proto bez připojených souborů.

4.8.3 Import do DSpace

Výsledná adresářová struktura po převodu exportoných dat z T-Series do XML je již připravena pro import do DSpace pomocí přiložených nástrojů DSpace. Import a export dat v DSpace je v současné době řešen pomocí tříd, které se spouštějí z příkazového řádku pomocí speciálního spouštěcího souboru *dsrun*, který zajišťuje bezpečný přístup do databáze DSpace. Při importu je nutné zadat několik potřebných údajů, jako jsou email uživatele, pod kterým se záznamy budou vkládat, identifikátor kolekce do které se budou záznamy vkládat a také jméno mapovacího souboru. Tento mapovací soubor po importu

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<dublin_core>
  <dcvalue element="title" qualifier="none"  language="en">Testing and statistical feedback</dcvalue>
  <dcvalue element="contributor" qualifier="author">Šaloun, Petr</dcvalue>
  <dcvalue element="language" qualifier="iso">en</dcvalue>
  <dcvalue element="contributor" qualifier="author">Šalounová, Dana</dcvalue>
  <dcvalue element="contributor" qualifier="author">Madryová, Anna</dcvalue>
  <dcvalue element="title" qualifier="alternative"  language="en">Testování a statistická zpětná vazba</dcvalue>
  <dcvalue element="identifier" qualifier="citation" >Sborník vědeckých prací Vysoké školy báňské - Technické univerzity
Ostrava. Řada elektrotechnická. 1999, roč. 5, č. 1, s. 23-32 : il.</dcvalue>
  <dcvalue element="relation" qualifier="ispartofseries">Sborník vědeckých prací Vysoké školy báňské - Technické
univerzity Ostrava. Řada elektrotechnická</dcvalue>
  <dcvalue element="publisher" qualifier="none">Vysoká škola báňská - Technická univerzita Ostrava</dcvalue>
  <dcvalue element="identifier" qualifier="issn">1210-048X</dcvalue>
  <dcvalue element="date" qualifier="issued">1999</dcvalue>
  <dcvalue element="type" qualifier="none">Článek</dcvalue>
</dublin_core>
```

Obrázek 9: Ukázka článku po konverzi v souboru dublin_core.xml

Obrázek 10: Ukázka článku v DSpace

obsahuje seznam identifikátorů všech importovaných záznamů a je možné jej použít pro odstranění naimportovaných záznamů nebo jejich modifikaci.

Import byl nejprve testován na několika (asi 600) záznamech. Když se zdály být výsledky konverze v pořádku, naimportovaly se všechny potřebné záznamy, kterých bylo asi 24200. Celý import trval na našem testovacím serveru asi 23 hodin, což je poměrně dlouhá doba, ale jelikož se takový rozsáhlý import provádí pouze jednou, není třeba chápat to jako chybu DSpace.

4.9 Testování

Aby bylo možné nasadit systém do reálného provozu a mohl poskytovat služby široké veřejnosti, byla potřeba jej řádně otestovat. Prvotní testování jsem prováděl sám a hledal jsem hlavně chyby v upraveném kódu. V dalším testování se zkoušely hlavně vkládací formuláře. S testováním vkládacích formulářů mi pomáhaly Mgr. Pavla Rygelová a Mgr. Alena Hausková. Testování probíhalo v iteračním cyklu, dokud se systém z pohledu knihovníků nejevil úplně v pořádku. Následovalo testování a kontrola importovaných záznamů a poté procházení a vyhledávání v těchto záznamech.

V poslední fázi jsme kontrolovali a upravovali správnost a vhodnost českého překladu uživatelského rozhraní DSpace. Některé výrazy byly sice přeloženy správně, ale nezapadaly do kontextu českého prostředí, a proto byly nahrazeny jinými.

4.10 Zálohování a přesun na nový server

Když bylo vše ve stavu, kdy systém mohl přejít do ostrého provozu, bylo nutné kompletní systém přenést na nový server. K tomu bylo třeba systém zálohovat a obnovit na druhém stroji. DSpace je možné zálohovat třemi způsoby, přičemž pro migraci fungujícího systému se nejvíce hodí druhý způsob. Zálohovat je možné těmito způsoby:

1. je možné zálohovat kompletně celý souborový systém nebo jen část nutnou pro přenos DSpace. Jsou to hlavně adresáře s instalací DSpace a systémové adresáře, ve kterých jsou uložena data PostgreSQL. Je možné také zálohovat soubory serveru Apache Tomcat a adresář se zdrojovými kódy DSpace pro případné další úpravy. Následná obnova dat se provede zkopírováním zálohovaných adresářů do nového systému,

2. je možné zálohovat databázi na programové úrovni a adresář s úložištěm souborů DSpace. K zálohování databáze slouží příkazy *pg_dump* a *pg_dumpall*, které zálohují obsah databáze jako sekvenci SQL příkazů do textového souboru. Tento soubor lze poté spustit na jiném počítači a SQL dotazy pro vytvoření a naplnění databáze se provedou. Při tomto způsobu obnovy databáze se musí spustit ještě soubor s SQL dotazy pro nastavení primárních klíčů, aby PostgreSQL nepřiřazoval již použité klíče. Adresář s úložištěm pak stačí pouze zkopírovat na správné místo a zbytek souborů aplikace se obnoví ze zdrojových kódů,

3. asi nejméně vhodná možnost pro přenos celého systému je export a následný import dat. DSpace umožňuje exportovat pouze záznamy z jednotlivých kolekcí, takže

bychom museli v novém systému nejdříve vytvořit příslušné kolekce a komunity a teprve potom přenést data jednotlivých kolekcí. Tato možnost je nejlepší pro částečnou migraci dat, jako je například převod kolekce z testovacího na produkční server.

4.11 Budoucí vývoj DSpace

Na DSpace se neustále pracuje a stále se do něj doplňují nové funkce a vylepšují se ty stávající. V našem případě jsme použili poslední verzi 1.3.2, která vyšla v říjnu 2005. V současné době se připravuje verze 1.4, zatím je dostupná k testování pouze beta verze. V této novější verzi by mělo být opraveno mnoho chyb a doplněno hodně nových funkcí. Z těch nejpodstatnějších změn to jsou:

- vylepšená práce se skupinami uživatelů, kde skupiny mohou obsahovat jiné skupiny,
- změny v autentizčním systému, lze nastavit jaká metoda autentikace (LDAP, heslo a jiné vlastní metody) se použije pro které uživatele,
- možnost definovat a používat více metadatových schémat, včetně vlastního vytvořeného,
- je možné procházení záznamů podle klíčových slov,
- možnost nastavit metadata, která se budou zobrazovat při výpisu záznamů. Ve výchozím nastavení se zobrazuje datum, název a autor a je možné přidat další.

Jako další cíle si vývojáři určili zlepšení podpory pro jiné jazyky, než je angličtina a možnosti jejich přepínání nezávisle na nastavení prohlížeče. Dále internacionalizace zbývajících částí systému (vkládací formuláře, emaily, nápověda) a vylepšení konfigurovatelnosti současných funkcí.

5 Popis příručky

Součástí zadání práce bylo vytvoření příručky k vybranému systému. Příručka by měla být vytvořena pomocí DocBook, aby z ní bylo možno vygenerovat několik výstupních formátů. Měla by sloužit knihovníkům a administrátorovi DSpace pro snadnější seznámení se se systémem a názornou ukázkou vysvětlit postupy provádění některých operací. V některých částech by měla nahradit originální anglickou dokumentaci a jinde ji jen doplnit. Uživatelská dokumentace vytvářena až na výjimku nebyla, protože práce se systémem DSpace je velice jednoduchá a navíc je v systému velice kvalitní nápověda, kterou by v případě potřeby stačilo přeložit do češtiny. Výjimkou je nápověda k vyhledávání záznamů, která byla vytvořena a přiložena k originální nápovědě a vysvětluje rozšířené vyhledávání.

Vytvořená dokumentace je dostupná na přiloženém CD, jehož obsah je popsán v příloze C. Krátkou ukázku z této dokumentace můžete vidět na obrázcích v příloze A. Kompletní dokumentace je dostupná jako sada XHTML stránek a jako samotný PDF soubor přichystaný k tisku. Díky použití DocBooku k tvorbě příručky je možné vytvořit i jiné formáty.

5.1 DocBook

DocBook [33] je formát založený na XML souborech pro zápis textových dokumentů. Původně byl vytvořen jako formát pro tvorbu dokumentací k softwaru, ale dnes se používá pro spoustu jiných typů dokumentů. Protože je DocBook založen na XML souborech, umožňuje oddělit vzhled dokumentů od jejich obsahu. XML jazyk byl podrobněji popsán v kapitole 4.1.7, takže jeho výhody není třeba popisovat. DocBook definuje sadu vlastních značek především pro tvorbu dokumentací, ale také sadu stylů, které umožňují generovat různé výstupní formáty.

Tvorba dokumentů v DocBooku má proti jiným metodám řadu výhod. Pokud vytváříme dokumentaci k rozsáhlejšímu systému, můžeme potřebovat části dokumentace vložit například do nápovědy samotné aplikace, do celkové příručky nebo kdekoliv jinde. XML soubor DocBooku může být rozdělen na několik částí a tyto části se mohou použít v mnoha případech pro generování požadovaných částí dokumentace v různých formátech. Pokud budeme potřebovat provést změnu v dokumentaci, není třeba opravovat několik samostatných souborů, ale opravíme dokumentaci pouze v jednom XML souboru a z něj opět vygenerujeme opravené verze dokumentace. Můžeme také potřebovat generovat několik verzí dokumentu lišící se pouze v několika málo kapitolách. V tomto případě stačí napsat jen odlišné kapitoly a zbytek dokumentu zůstane pouze v jedné kopii. Vkládáme-li do textu obrázky, docbook umožňuje vložení odkazů na více typů obrázků, například vektorové pro tisk nebo bitmapové pro prezentaci na monitoru. Při generování výstupního souboru se použije vhodnější typ obrázku. Implementace DocBooku je nezávislá na použité platformě, takže jej lze použít na různých operačních systémech. Nezanedbatelnou výhodou je také cena, protože DocBook je poskytnut zdarma.

Na následující ukázce je stručný popis jednoduché knihy, popsané v DocBooku. Soubor má klasickou XML strukturu, jen za hlavičkou XML se zapisuje ještě definice typu

dokumentu pro DocBook. V následujícím případě jde o DocBook verze 4.2. Dále následují značky DocBooku definované pro tvorbu knihy. Atributem *lang* je nastaven jazyk knihy na český, aby se v knize generovaly české názvy kapitol (kapitola, obsah a další). Následují informace o knize (název, údaje o autorovi) a dále již samotný úvod a kapitoly knihy.

```
<?xml version='1.0' encoding='iso-8859-2'?>
<!DOCTYPE book PUBLIC '-//OASIS//DTD DocBook XML V4.2//EN'
  'http://www.oasis-open.org/docbook/xml/4.2/docbookx.dtd'>
<book lang="cs">
  <bookinfo>
    <title>Moje kniha</title>
    <author>
      <firstname>Dušan</firstname>
      <surname>Jalůvka</surname>
    </author>
  </bookinfo>
  <preface>
    <title>Úvod</title>
    <para>První odstavec úvodu.</para>
    <para>Druhý odstavec úvodu.</para>
  </preface>
  <chapter>
    <title>První kapitola</title>
    <para>Text první kapitoly</para>
  </chapter>
</book>
```

Z takového souboru může být vygenerován text knihy nebo v závislosti na použitém stylu i s obsahem a titulní stranou. Styly definují, jak se má dokument zobrazit nebo vytisknout a jsou zapisovány stylovými jazyky. V současné době se používají pro zápis stylů XSL a DSSSL jazyky, které umožňují převod do mnoha různých formátů jako jsou RTF, PDF, PostScript, HTML, XHTML, HTML Help a další. U formátů HTML a XHTML si můžeme zvolit, zda se vygeneruje jeden velký soubor nebo se text rozdělí podle kapitol do několika menších souborů. Pokud máme na výstupní formát zvláštní požadavky, můžeme použít připravený styl a doplnit je do něj.

Spojení vytvořeného XML dokumentu s některým formátovacím stylem zajišťuje stylový procesor, jehož výstupem je soubor nebo sada souborů v požadovaném formátu. Mezi nejpoužívanější formátovací procesory patří *xsltproc*, *Saxon* nebo *Jade*.

5.2 Administrátorská dokumentace

Administrátorská dokumentace má sloužit správci serveru a administrátorovi DSpace, aby byli schopni zprovoznit DSpace ze zdrojových kódů a provést na něm potřebné úpravy a nastavení. Práce popsané v této příručce vyžadují alespoň základní znalosti

správy operačního systému GNU/Linux a alespoň základní znalosti platformy Java. Správce systému a serveru by měl být schopen pracovat i s databází PostgreSQL. Měl by také porozumět systému DSpace, aby byl schopen zprovoznit systém, pokud se vyskytne nějaká chyba. Protože se při práci s DSpace předpokládá i práce s textovými nástroji na straně serveru, měl by být administrátor částečně knihovníkem nebo s ním alespoň spolupracovat.

V administrátorské příručce jsou podrobně popsány všechny problémy, se kterými jsem se při instalaci a administraci DSpace setkal. Jsou to například tyto úkony:

- Na začátku je popsána příprava na instalaci, kde jsou popsány všechny softwarové balíky, které je třeba mít před instalací nainstalované. Instalace těchto balíků není rozepsána podrobně, jen jsou zmíněny kroky, které mají vliv na výsledné chování DSpace a které by se neměly opomenout. Jsou to například operační systém, platforma Java, databázový systém PostgreSQL a servlet kontejnér Apache Tomcat.
- Dále jsou popsány všechny provedené změny na originálních zdrojových kódech a popsán postup začlenění změn přepsáním upravených souborů.
- Instalace je popsána celkem podrobně krok za krokem, takže by neměl být problém nainstalovat DSpace ze zdrojových kódů.
- Část dokumentace je věnována zálohování, protože zálohování je nezbytná součást každého systému a navíc je potřebné při převodu systému na jiný server.
- Aby bylo možné provést změny ve vkládacích formulářích, je v dokumentaci popsána úprava formulářů v souboru *dublin_core.xml*.
- Také jsou tam popsány základní nastavení, která se provádějí v souboru *dspace.cfg*.
- Stejně tak je tady popsán postup převodu dat ze systému T-Series a následný import do DSpace. Možnosti exportu dat z kolekcí DSpace jsou jen zmíněny a je zde odkaz na originální dokumentaci.

5.3 Knihovnická dokumentace

V knihovnické dokumentaci jsou popsány úkony, které budou provádět správcové kolekce nebo pracovníci knihovny prostřednictvím webového administračního rozhraní nebo pomocí rozhraní schvalovacích a dohlížecích procesů. Při vysvětlování postupu provádění knihovnických úkonů je použito názorných ukázek, které zobrazují grafické uživatelské rozhraní. K těmto ukázkám je vždy popsán případ, kdy se taková operace provádí a jaké následky může mít na záznamy v archívu. Knihovnická dokumentace vysvětluje nejpoužívanější funkce a doplňuje originální nápovědu zabudovanou v DSpace. Tato nápověda je psána anglicky v XHTML a v případě potřeby není problém ji lokalizovat do češtiny. Já jsem to při lokalizaci DSpace neudělal, protože se u uživatelů DSpace předpokládá alespoň základní znalost angličtiny.

V knihovnické dokumentaci jsou popsány následující operace, které provádí buď správce kolekce nebo pověřená osoba, která má zkontrolovat, přijmout nebo editovat záznam před zařazením do hlavního archívu.

- Je posáno vytváření kolekcí a komunit, popsány možnosti při vytváření a pravidla pro členění komunit a podkomunit.
- Dále je popsána správa těchto kolekcí a komunit, popsány oprávnění administrátora kolekce a možnosti mapování záznamů mezi kolekcemi.
- Nastavení přístupových politik pro komunity, kolekce, záznamy a soubory je věnována samostatná kapitola. Je podrobně vysvětlen význam jednotlivých nastavení a ty jsou demonstrovány na příkladech.
- Dokumentace se věnuje také nastavení řízenému koloběhu dokumentů, tzv. workflow procesu. Jsou popsány možnosti nastavení schvalování a editace metadat. Je zde také popsán průběh při schvalování dokumentů a zařazení do hlavního archívu.
- DSpace poskytuje možnost dohlížení vedoucích nad pracemi studentů nebo možnost nastavení spolupráce na jednom záznamu. V dokumentaci je popsáno nastavování těchto dohledů a možnosti dohlížejících uživatelů.
- Jedna kapitola je věnována také editaci a mazání záznamů. V této kapitole jsou popsány možnosti a rizika při mazání a editaci.

6 Závěr

Hlavním cílem této práce bylo poskytnout Ústřední knihovně Vysoké školy báňské – Technické univerzity Ostrava systém úložiště digitálních dat a do tohoto systému převést naskenované články ze sborníků vědeckých prací a údaje o kvalifikačních pracích. Tento úkol byl splněn a všechny digitální dokumenty nebo případá metadata dokumentů jsou již importovány v systému DSpace. Pro systém DSpace jsme se rozhodli po porovnání se systémem Eprints. Po důkladném testování pracovnicemi knihovny je DSpace nyní připraven na ostrý provoz a může být převeden na hlavní server knihovny. Mimo hlavní obsah byl DSpace rozšířen jako univerzální repozitář, takže bude nabídnut univerzitě k volnému užití. To znamená, že si může kdokoliv z univerzity požádat o vytvoření své kolekce a prezentovat v DSpace své dokumenty jako například výukové materiály, skripta, publikace a další.

Při provádění úprav a při testování bylo nalezeno několik dalších možných návrhů, které by zjednodušily práci s DSpace a umožnily tak další rozšíření obsahu. Tyto návrhy mohou být do budoucna zpracovány a implementovány do DSpace. Mezi takové návrhy například patří kontrola jedinečnosti identifikátorů jako je signatura nebo přírůstkové číslo, kdy by DSpace automaticky novým záznamům přiřazoval tyto identifikátory na základě stanovených pravidel tak, jak již přiřazuje handle identifikátory. Dalším možným vylepšením by byla úprava rozšířeného vyhledávání, kde by se místo kódů jazyka vybíral jazyk z předdefinovaného seznamu.

Při této práci jsem získal řadu cenných zkušeností a znalostí a oblíbil jsem si systém DSpace jako takový. Navíc se mi zamlouvá myšlenka svobodného softwaru, takže je možné, že některé z výše zmíněných funkcí doplním v rámci svého volného času a nadále budu s knihovnou spolupracovat.

7 Literatura

[1] NEMETH, Evi; SNYDER, Garth; HEIN, Trent R. *LINUX : Kompletní příručka administrátora*. Brno : Computer Press, 2004. xxxiii, 828 s. ISBN 80-7226-919-4.

[2] TANSLEY, Robert; STUVE, David; BASS, Mick. *DSpace System Documentation* [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://dspace.org/technology/system-docs/index.html>.

[3] KREJČÍŘ, Vlastimil *Univerzální digitální repozitář : diplomová práce* [online]. Brno : Masarykova univerzita, 2005 [cit. 2006-04-27]. 116 s. Dostupný na WWW: <http://eprints.rclis.org/archive/00005076/>.

[4] VITÁSEK, Jan. *WWW prezentace sborníku vědeckých prací : diplomová práce*. Ostrava : Vysoká škola báňská – Technická univerzita Ostrava, 2003. 48 s.

[5] PASTUSZEK, Michal. *Úložiště digitálních dat pro potřeby ÚK VŠB–TU Ostrava II : diplomová práce*. Ostrava : Vysoká škola báňská – Technická univerzita Ostrava, 2006.

[6] BARTOŠEK, Miroslav. *Digitální knihovny* [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://www.ics.muni.cz/mba/dl-datakon01.pdf>.

[7] VANNEVAR, Bush. *As We May Think*. Atlantic Monthly, 1945.

[8] LICKLIDER, J. C. R. *Libraries of the Future*. Cambridge : The MIT Press, 1965.

[9] ŽABIČKA, Petr. *OAI–PMH : Protokol pro metadatovou interoperabilitu* [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://knihovny.cvut.cz/akp2003/sbornik/05_zabicka.pdf>.

[10] BARTOŠEK, Miroslav. *Digitální knihovny : Teorie a praxe* [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://eprints.rclis.org/archive/00005061/01/DL-Bartosek-final2.pdf>.

[11] DSpace [online]. MIT, 2006 [cit. 2006-04-27]. Dostupné na WWW: <http://dublincore.org>.

[12] Dublin Core Metadata Initiative [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://dublincore.org>.

[13] Metadata Object Description Schema [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://www.loc.gov/standards/mods>.

[14] Metadata Encoding & Transmission Standard [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://www.loc.gov/standards/mets>.

[15] Resource Description Framework [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://www.w3.org/Metadata/Activity.html>.

[16] Open Archives Initiative. *Protocol for Metadata Harvesting* [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[17] Z39.50 [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.loc.gov/z3950/agency/>.

[18] OpenURL [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://library.caltech.edu/openurl/>.

[19] CNRI Handles [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.handle.net/>.

[20] Digital Object Identifier [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.doi.org/>.

[21] E-LIS Digital library [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://eprints.rclis.org/>.

[22] Berkley Software Distribution License [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.opensource.org/licenses/bsd-license.php>.

[23] Storage Request Broker [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.sdsc.edu/srb/>.

[24] Lucene Search Index [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://lucene.apache.org/>.

[25] Creative Commons License [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://creativecommons.org/>.

[26] Sun Java Technology [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://java.sun.com/>.

[27] Java Server Pages [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://java.sun.com/products/jsp/>.

[28] PostgreSQL database [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.postgresql.org/>.

[29] Apache Tomcat [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://tomcat.apache.org/>.

[30] Open LDAP [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.openldap.org/>.

[31] Extensible Markup Language (XML) [online]. [cit. 2006-04-27]. Dostupné na WWW:
<http://www.w3.org/XML/>.

[32] VeriSign certifikační autorita [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://www.verisign.com/>.

[33] DocBook [online]. [cit. 2006-04-27]. Dostupné na WWW: <http://www.docbook.org/>.

A Ukázka vytvořené příručky

Na následujících stránkách jsou ukázky z vytvořené příručky administrátora a knihovníka, které byly z DocBooku vygenerovány do formátu PDF.

První stránka dokumentace

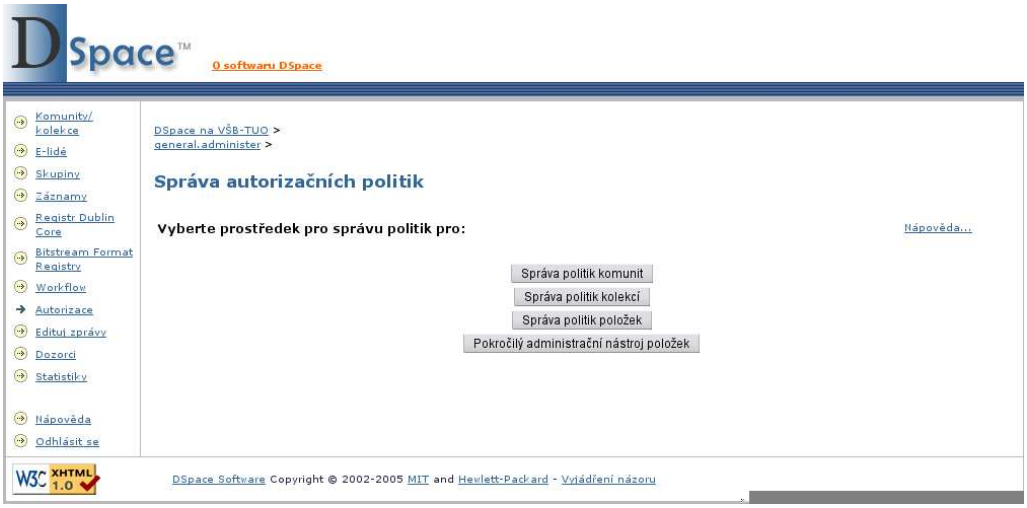
Druhá stránka dokumentace

Třetí stránka dokumentace

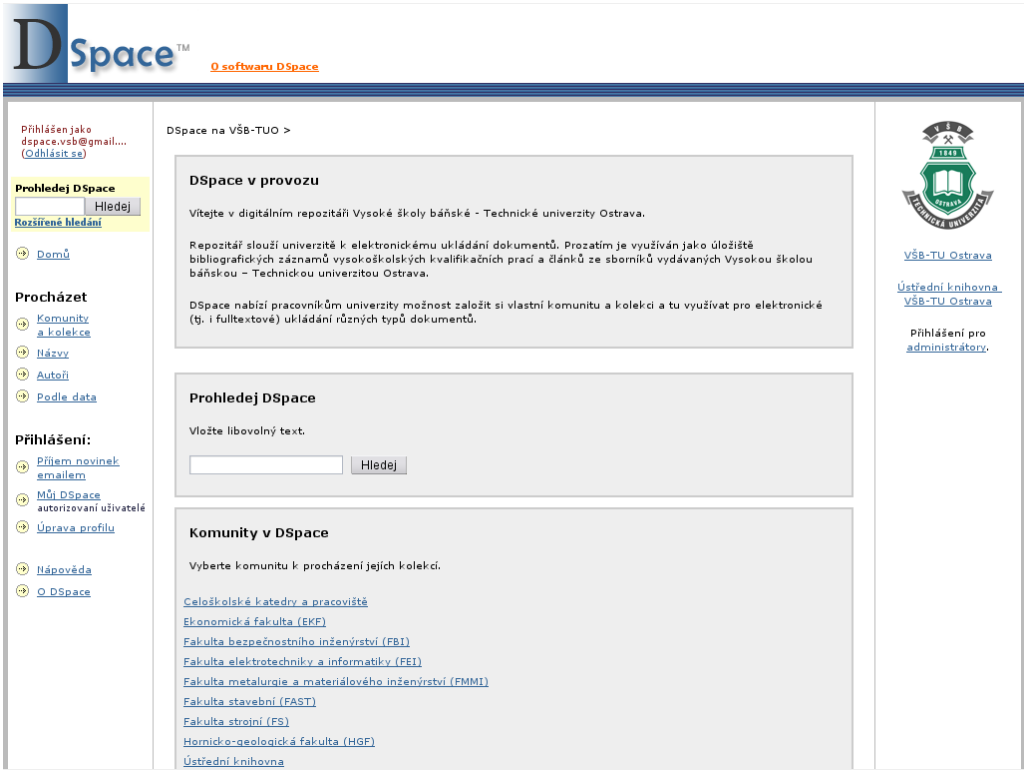
Čtvrtá stránka dokumentace

B Ukázka uživatelského rozhraní

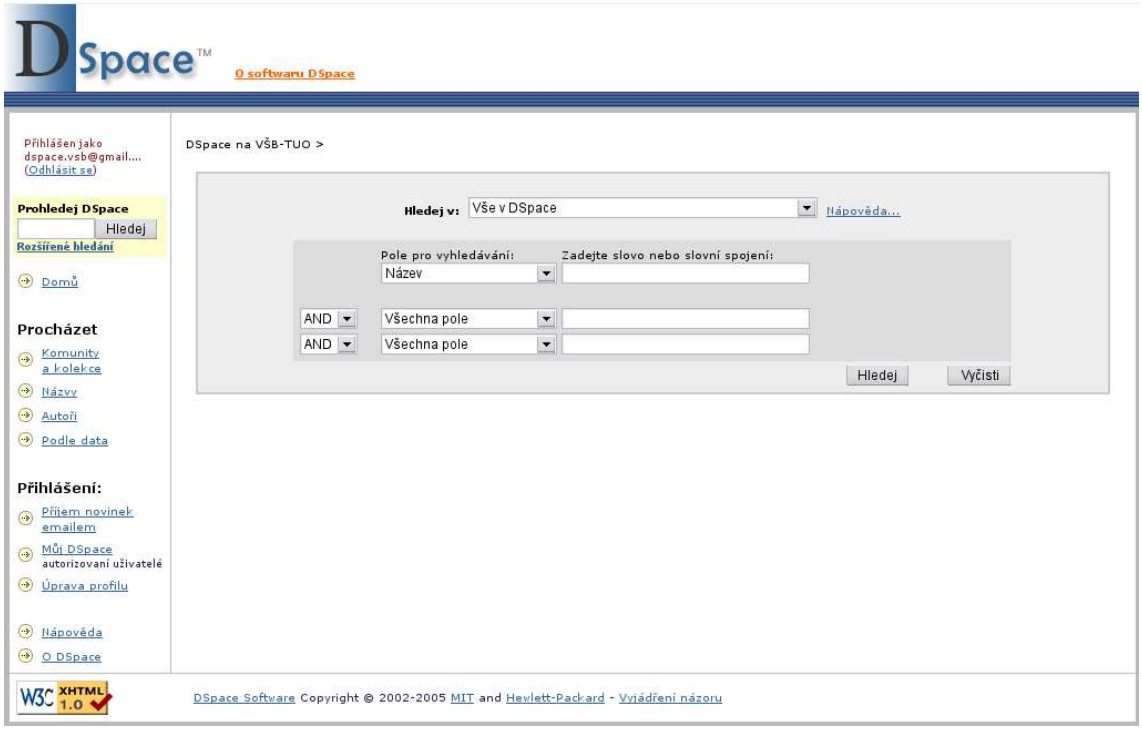
Nyní následuje několik obrázků s náhledy na grafické uživatelské rozhraní DSpace.



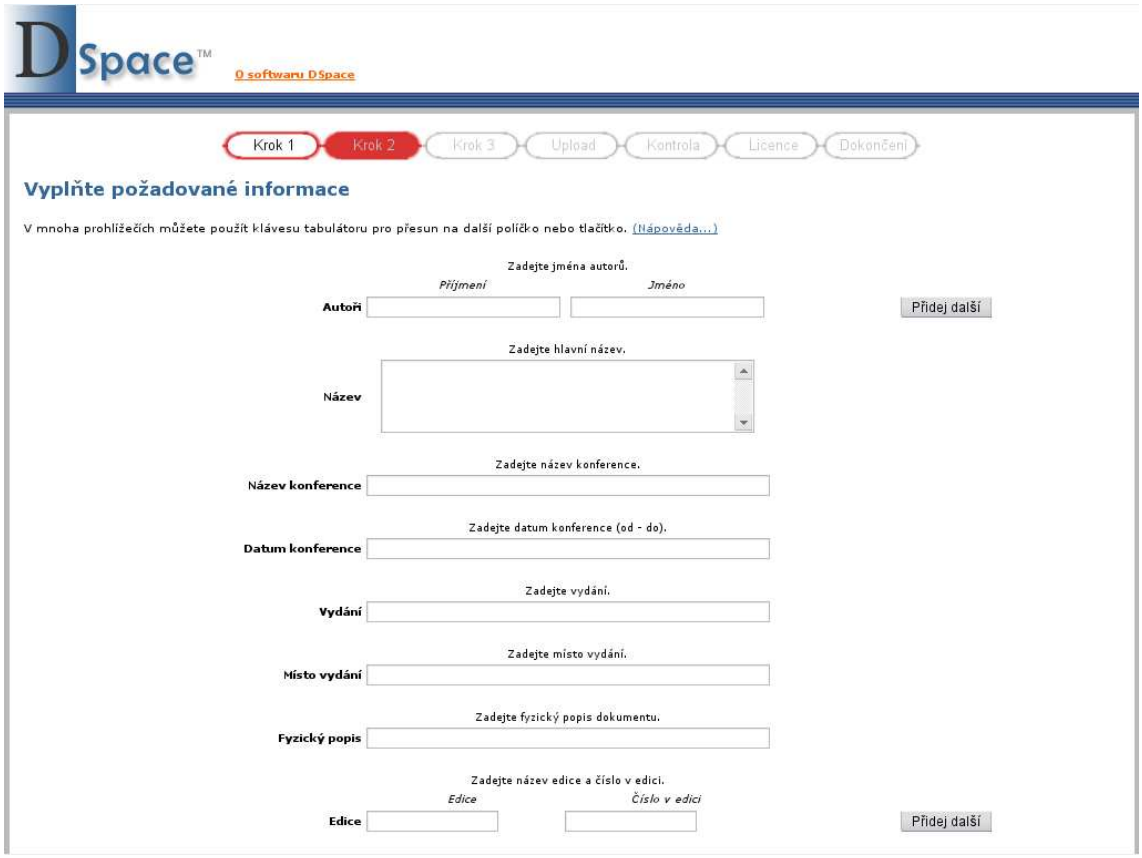
Obrázek 11: Administrátorské menu



Obrázek 12: Úvodní stránka DSpace



Obrázek 13: Rozšířené vyhledávání



Obrázek 14: Vkládací formulář

C Obsah přiloženého CD

Na přiloženém CD jsou všechny elektronické materiály, které vznikly při tvorbě této práce a při instalaci a úpravách DSpace. CD má následující strukturu:

- **dspace-source** – kompletní zdrojové kódy DSpace se všemi provedenými úpravami,
- **konverze** – program pro převod formátu dat T-Series do XML Dublin Core včetně dokumentace k programu,
- **prirucka** – knihovnická a administrátorská dokumentace v DocBooku, PDF a jako sada XHTML stránek,
- **software** – instalační archívy softwaru potřebného pro instalaci DSpace (Apache Tomcat, PostgreSQL, originál dspace–source),
- **text** – elektronická verze textu této práce v systému LaTeX a vygenerovaném formátu PDF,
- **upravy** – adresář s jednotlivými úpravami,
  - **config** – soubory s nastavením Apache Tomcat a DSpace,
  - **forms** – soubory s upravenými formuláři,
  - **jsp** – upravené JSP stránky grafického rozhraní DSpace,
  - **lokalizace** – lokalizované soubory s texty DSpace v několika kódováních,
  - **registries** – soubory s inicializačními nastaveními registru Dublin Core.