T A R T U Ü L I K O O L
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika eriala

Marten Teino

Uurimisrühmasisene teaduskirjanduse andmebaas

Bakalaureusetöö (10 ap)

Juhendajad: Jaak Vilo, PhD
Hedi Peterson, MSc

Autor: ..... "....." mai 2007
Juhendaja: ..... "....." mai 2007
Õppetooli juhataja: ..... "....." 2007

TARTU 2007

Sisukord

Sissejuhatus

Teaduskirjandusel, mida iga päev suurel hulgal produtseeritakse ja publitseeritakse, on oluline roll oma uurimisvaldkonnas toimuvaga kursis olemiseks. Digitaalsel kujul kirjanduse säilitamiseks ja levitamiseks on olemas mitmeid andmebaase – Pubmed[1] [WBB+06], Citeseer[2] [LCB+06], ACM Digital Library[3] [Whi01].

Uurimisrühmad ja teadusgrupid kasutavad oma töös erinevatest allikatest pärinevaid teadusartikleid. Erinevate allikate rohkus ja pakutava kirjanduse suur hulk raskendab olulise leidmist. Relevantset informatsiooni on vaja kuidagi uurimisrühmasiseselt talletada nii, et see vajadusel kiiresti leitav oleks. Tihti on ka oluline teadusgrupi kõikide liikmete ligipääs.

Uurimisgruppide töö tulemusel luuakse ka mitmesuguseid materjale nagu ettekanded, raportid, presentatsioonid, postrid, artiklite mustandid jne. Sarnaselt teaduskirjandusega on ka see materjal uurimisrühma kontekstis oluline ja tuleb kuskil säilitada ning tagada kättesaadavus.

Tartu Ülikooli, Egeeni ja Eesti Biokeskuse bioinformaatikuid ühendav teadusgrupp BIIT (Bioinformatics, Algorithmics, and Data Mining group) tegutseb dr. Jaak Vilo juhtimisel eesmärgiga rakendada informaatikaalaseid teadmisi, oskusi ja vahendeid bioloogia ning geenitehnoloogia probleemide lahendamisel [BII]. Teadusgrupi kasvades on tekkinud tungiv vajadus kirjanduse andmebaasi järele, kus saaks säilitada loetavat teaduskirjandust ja muud materjali. Sellest tulenevalt on käesoleva töö peamiseks eesmärgiks täpsustada ja analüüsida teadusgrupi BIIT vajadusi ning luua keskkond kirjanduse talletamiseks.

[1] http://www.ncbi.nlm.nih.gov/entrez
[2] http://citeseer.ist.psu.edu
[3] http://portal.acm.org/dl.cfm

Käesoleva töö esimeses osas kirjeldan ja analüüsin esmalt uurimisrühma BIIT vajadusi. Kuna digitaalsel kujul informatsiooni säilitamise probleem pole maailmas esmakordne, tundus mõistlik uurida, millised süsteemid on selleks juba olemas. Teises peatükis võrdlengi kolme digitaalse materjali repositooriumi tarkvara – Greenstone'i, DSpace'i ja Fedora funktsionaalsust ja toon ära nende eelised ning puudused.

Kolmandas peatükis keskendun kolmest vaatluse all olnud süsteemist sobivaimaks osutunud DSpace'i funktsionaalsuse põhjalikumale kirjeldamisele.

Neljandas peatükis käsitlen antud töö mahukaimat osa – DSpace'i kohandamist teadusgrupi BIIT vajadustele ning lisafunktsionaalsuse realiseerimist. Seejärel annan ka ülevaate DSpace'i kasutamisaktiivsusest uurimisrühma BIIT liikmete poolt ning toon ära pisut statistikat.

Peatükk 2

Uurimisrühma BIIT vajaduste analüüs

Oma igapäevatöös kasutavad teadusgrupi BIIT liikmed erinevatest allikatest pärit teaduskirjandust, milleks on enamjaolt PDF (Portable Document Format) või PS (Postscript) formaadis failid. Seni on uurimisgrupi liikmed säilitanud kirjandust oma isiklikes arvutites. Arvutis failidena salvestatud materjalist oli aga vajaliku leidmine tihtipeale aeganõudev, sest sooritada ei saa täistekstiotsingut, leidmaks kiiresti faile, milles esineb näiteks fraas "geenide avaldumine" või "DNA" või ka mõlemad korraga.

Erinevates arvutites kirjanduse talletamise tulemusena puudus teadusgrupi liikmetel selge ülevaade sellest, milliseid artikleid rühmasiseselt kasutatud on ja kus need täpselt asuvad. Ülevaate puudumise tõttu ei osatud ka otsida. Materjali asukoha teadmisel oli järgmiseks probleemiks, kuidas vastav fail kätte saada, kui see asub näiteks teise inimese arvutis. Sellistel puhkudel toimus levitamine enamasti e-kirja teel, mis võis olla aeganõudev.

Metaandmeteks või *metainfoks* nimetatakse andmeid, mis kirjeldavad andmeid [BDM03]. Artiklit, raamatut, ettekannet vms kirjeldavaks informatsiooniks on näiteks autor, pealkiri, publitseerimise kuupäev, sisukokkuvõte jne. Failisüsteemis säilitatavatele failidele sellist informatsiooni lisada ei saa, metainfost otsimisest rääkimata.

Publitseeritava teaduskirjanduse hulk on suur, selle kvaliteet ja tähtsus aga erinev. Nii mõnigi artikkel on juba teadaolevate faktide korduv kirjapa-

nemine ning seetõttu uut ja huvitavat ei sisalda. Ajaraiskamise vältimiseks on mõistlik lugeda vaid "olulist" kirjandust. Oluliste artiklite tuvastamiseks on mitmeid võimalusi nagu näiteks hindamine ja kommenteerimine lugejate poolt. Palju loetud ja kommenteeritud artikkel on tõenäoliselt olulisem kui artikkel, mille vastu pole huvi tuntud. Tuleb märkida, et failisüsteemis säilitatavaid faile pole võimalik ei kommenteerida ega hinnata.

Mainitud probleemide lahendamiseks oli vaja luua kirjanduse andmebaas, kuhu saaks koondada kogu teadusgrupi liikmete poolt loetava ja ka produtseeritava materjali. Interneti laiast levikust tingituna oli iseenesestmõistetav, et loodav keskkond peab olema andmebaasiga veebirakendus.

Arvestades asjaoluga, et seni arvutites säilitatud materjali läheb ka tulevikus vaja, pidi seda saama kirjanduse andmebaasi üle viia. Failide hulka hinnates, oli selge, et nende ükshaaval käsitsi lisamine pole mõeldav. Importimine pidi seega olema võimalikult lihtne ja kiire.

Iga artikli, raamatu, ettekande kohta peab saama kirjanduse andmebaasis hoida ka metaandmeid. Lihtsaim lähenemine olnuks keskkonna kasutajate ehk teadusgrupi BIIT liikmete poolt metainfo sisestamine. See polnud just hea lahendus, sest käsitsi andmete sisestamine on aeganõudev ja seetõttu tülikas. Olemasoleva kirjanduse päritolu analüüsides selgus, et enamik pärineb suurtest teaduskirjanduse andmebaasidest nagu Pubmed, Pubmed Central [1] või Biomed Central[2]. Nendes keskkondades on olemas paljude publitseeritud teadusartiklite metainfo, mida on ehk võimalik automaatselt pärida ning uurimisrühma kirjanduse andmebaasi salvestada.

Eeldades, et loodavasse keskkonda koguneb aja jooksul tuhandeid faile, tundus ülevaatlikkuse tagamise eesmärgil mõistlik materjali grupeerimise ja klassifitseerimise nõue. Grupeerida ja klassifitseerida on võimalik mitmeti – temaatika (nt DNA, RNA, geenide avaldumine), tüübi (nt raamatud, teadusajakirjad, ettekanded, juhendid), publitseerimise aja või muude kriteeriumite järgi. Võimaldamaks vajalikku kiiresti ja vähese vaevaga leida, oli tähtis ka otsingumootor, mille abil otsida nii täistekstist kui ka metaandmetest. Otsingutulemuste hulga kitsendamiseks annab hea võimaluse grupeerimise kombineerimine otsinguga. Selle all tuleb mõista otsimist ühest konkreetsest

[1] http://www.pubmedcentral.nih.gov

[2] http://www.biomedcentral.com

grupist.

Olulise kirjanduse tuvastamiseks pidi loodaval keskkonnal kindlasti olema võimalus materjali kommenteerida ja hinnata. Hinnete põhjal arvutatud skoori võiks kasutada näiteks otsingutes tulemuste sorteerimisel. Oluline on ka kasutajate tegevuse jälgimine eesmärgiga tuvastada huvitavaid artikleid. Huvitavate artiklite kohta võiks kasutajatele anda soovitusi.

Kokkuvõtvalt pidid teadusgrupi BIIT kirjanduse andmebaasil olema järgmised võimalused:

1. üksiku artikli ja ka suure hulga kirjanduse (failide) võimalikult lihtne ning kiire lisamine
2. metainfo automaatne lisamine
3. nii metaandmetest kui ka täistekstist otsimist võimaldav otsingumootor
4. materjali grupeerimine ja klassifitseerimine
5. kirjanduse kommenteerimine ja hindamine
6. soovituste andmine

Peatükk 3

Digitaalsete dokumentide repositooriumid

Repositooriumi all mõistetakse juurdepääsetavate inforessursside kogumit, mis tavaliselt koosneb ühise otsingumootoriga andmebaasidest. Oluline on, et talletada saab igasugust digitaalsel kujul materjali: teksti, pilte, video- ja helifaile jne.

Internetist olemasolevaid süsteeme otsides, valisin välja kolm vabavaralist avatud lähtekoodiga repositooriumi tarkvara – Greenstone'i[1] [WBBM00], DSpace'i[2] [BB01] ja Fedora[3] [KSCS04].

3.1 Võrdlemise kriteeriumid

Tuginedes peamiselt Yan Han'i artiklile [Han04] toon alljärgnevalt ära digitaalsete repositooriumite võrdlemise kriteeriumid.

3.1.1 Algupärasuse säilimise tagamine

Sõltumata sellest, kuidas repositooriumis faile hoitakse, tuleb säilitada faili nimi ja suurus. Veendumaks faili korrektses lisamises, on vaja rakendada terviklikkuse kontrolli. Üks levinumaid meetodeid terviklikkuse kontrolliks

[1] www.greenstone.org
[2] www.dspace.org
[3] www.fedora.info

põhineb krüptograafiliste räsifunktsioonide MD5 (Message-Digest algorithm 5) või SHA1 (Secure Hash Algorithm 1) kontrollsumma arvutamisel ja võrdlemisel [SWZ05]. Kui räsifunktsiooni poolt tagastatav kontrollsumma on failil sama nii enne kui ka pärast lisamist, toimus lisamine korrektselt.

3.1.2 Metainfo standardid

Digitaalse materjali repositooriumis kasutatakse sisu kirjeldamiseks vähemalt ühte metainfo standardit. Mõned näited standarditest on DC[4] (Dublin Core), METS[5] (Metadata Enconding & Transmission Standard), MARC[6] (Machine-readable Cataloging).

Rohkem kui ühe standardi toetus on positiivne omadus, lihtsustades digitaalse materjali jagamist ja levitamist erinevate repositooriumite vahel [Dub03]. Importides näiteks Dublin Core standardile vastava metainfoga materjali, peab repositoorium seda standardit toetama. Vastasel juhul tuleb metaandmed käsitsi sisestada või näha vaeva metainfo konverteerimisega.

3.1.3 Informatsiooni kättesaadavus

Informatsiooni kättesaadavuse all mõistetakse kahte aspekti:

- püsivate identifikaatorite kasutamine
- välistele süsteemidele ligipääsu tagamine

Üsna tihti kasutatakse viitamiseks URL-e (Uniform Resource Locator). URL on otseselt seotud mingi ressursi (veebileht, pilt jne) füüsilise asukohaga, mille muutudes läheb viit "katki". Selle vältimiseks on mõeldud püsivad identifikaatorid (ik *persistent identifiers*). Aja jooksul on loodud ja rakendustes kasutust leidnud erinevad püsivate identifikaatorite realisatsioonid nagu DOI (Digital Object Identifier), CRNI *Handles* (Corporation for National Research Initiatives Handles), ARK (Archival Resource Keys), PURL (Persistent Uniform Resource Locators), URN (Uniform Resource Name) [HK06].

[4] http://dublincore.org
[5] http://www.loc.gov/standards/mets
[6] `http://www.loc.gov/marc`

Püsivate identifikaatorite põhiidee näeb ette globaalsete registrite kasutamist ja haldamist. Mingile ressursile, olgu selleks näiteks fail, määratakse globaalselt unikaalne identifikaator, mis registrisse kantakse ning faili füüsilise asukohaga seostatakse. Päringud tehakse registrisse, mis identifikaatorile vastava faili tagastab. Juhul, kui ressursi füüsiline asukoht peaks muutuma, tuleb vaid korrigeerida registrit ning kõik toimib endiselt.

Välistele süsteemidele ligipääsu võimaldamiseks on mõeldud näiteks klient-server standard OAI-PMH[7] (Open Archives Initiative Protocol for Metadata Harvesting). OAI-PMH defineerib üheselt reeglid üle HTTP (Hypertext Transfer Protocol) protokolli metainfo pärimiseks. Päringu vastuseks on kindlal kujul XML.

3.1.4 Süsteemi omadused ja kasutatavus

Antud kriteeriumi all vaadeldakse võrreldavate süsteemide omadusi ja kasutatavust. Näiteks milliseid võimalusi pakub otsingumootor, kas on võimalik sooritada täistekstiotsinguid? Milline on kasutajaliides ja kuidas on lahendatud repositooriumi sisu lehitsemine? Kas materjali lisamine toimub defineeritud töövoogu järgides? Kui paindlik on kasutajate ja õiguste haldus?

3.2 Greenstone

Greenstone'i digitaalse materjali säilitamise ja haldamise tarkvara esimene versioon valmis Waikato Ülikooli arvutiteadlaste töö tulemusel aastal 2000 [GRE]. Viimane uuendus versiooni 3.02 näol võeti kasutusele 2007. aasta veebruaris. Tarkvara on lubatud vabalt kasutatada GNU GPL (General Public Licence) tingimusi järgides. Programmeerimiskeeltena on kasutatud C++, Javat, Perli ning andmebaasimootorina GDBM-i (*Gnu Database Manager*).

Algupärasuse säilitamise nõuet täidetakse osaliselt – talletatakse küll faili suurus, kuid mitte selle esialgset nime. Faili terviklikkuse kontroll puudub, mistõttu pole võimalik tuvastada, kas lisamine toimus korrektselt.

[7] http://www.openarchives.org/OAI/openarchivesprotocol.html

Greenstone'is saab kasutada erinevaid metaandmete formaate. Aktsepteeritavad on nii Dublin Core, MARC kui ka METS.

Püsivaid identifikaatoreid ei kasutata, mistõttu viitamine pole usaldusväärne. Väliste süsteemidega suhtlemise võimaldamiseks kasutab Greenstone OAI-PMH standardit.

Repositooriumi kasutajad jagunevad kolme eeldefineeritud rolli – tavakasutajad, sisuhaldurid ning administraatorid. Iga roll määrab õiguste komplekti, mis lubavad sooritada tegevusi. Puuduseks on asjaolu, et rolle lisada ei saa.

Otsida saab täistekstist, kasutades nii metamärke kui ka tõeväärtusloogikat. Otsingutingimusi saab kombineerida konjunktsiooni, disjunktsiooni ning eituse abil. Repositooriumis talletatavaid dokumente saab grupeerida defineerides *kogusid* (ik *collection*) ja *alamkogusid* (ik *subcollection*).

Greenstone on mitmekeelne. Lisaks sellele, et infot hoitakse Unicode kodeeringus, on ka kasutajaliides erinevatesse keeltesse tõlgitav.

Seadistatava töövooprotsessi puudumine on antud süsteemi üheks miinuseks.

3.3 DSpace

Repositooriumitarkvara DSpace on mõeldud digitaalsel kujul informatsiooni säilitamiseks, indekseerimiseks ning kättesaadavaks tegemiseks. Tarkvara loomist alustati Massachusettsi Tehnoloogiainstituudi Raamatukogu (Massachusetts Institute of Technology Libraries) ja HP (Hewlett Packard) poolt 2000. aastal. Esimene ametlik versioon valmis 2002. aasta lõpus [DSP], viimane täiendus 1.4.1 tehti kättesaadavaks detsembris 2006. DSpace'i kasutamistingimused on määratud BSD litsentsiga. DSpace on programmeeritud Javas, andmebaasiks sobib nii PostgreSQL kui ka Oracle.

Algupärasuse säilitamise nõude täitmiseks talletatakse fail esialgse nimega, samuti salvestatakse selle suurus. Terviklikkuse kontroll põhineb MD5 räsifunktsiooni kasutamisel. Pärast faili üleslaadimist arvutatakse selle kontrollsumma, mida tuleb võrrelda esialgse faili kontrollsummaga, et veenduda lisamise korrektsuses.

Vaikimisi metainfo formaadiks on Dublin Core, samas võib defineerida ja

kasutada ka teisi formaate. Ühe piiranguna tuleb mainida asjaolu, et ühele dokumendile saab lisada metaandmeid vaid ühes formaadis. See tuleneb andmebaasi üks-ühesest seosest dokumendi ja metainfo formaadi vahel.

CRNI *Handles* [CRN] realisatsiooni kasutamine püsivate identifikaatoritena tagab repositooriumis sisu kättesaadavuse ka selle asukoha muutumisel. Välised süsteemid saavad metainfot pärida OAI-PMH standardile vastavate päringute abil.

Kasutajate ja õiguste haldus on üsna paindlik. Defineerida saab rolle ning neid kasutajatele omistada. Vaikimisi on olemas kaks rolli: administraatorid ning anonüümsed kasutajad.

Päringusüsteemi jaoks kasutatakse Javas realiseeritud Lucene [LUC] otsimootorit, mis võimaldab sarnaselt Greenstone'iga kasutada metamärke ja tõeväärtusloogikat.

Repositooriumis talletatavaid dokumente saab grupeerida kasutades kahte tasandit – *kommuun* (ik *community*) ja *kogu* (ik *collection*). Kommuuni all mõistetakse näiteks teaduskonda, instituuti või osakonda. Kogu saab defineerida mingi konkreetse valdkonna dokumentide tarvis. Igale kogule on muuhulgas võimalik luua *alamkogusid* (ik *subcollection*).

DSpace'il on kasutusel defineeritud töövooprotsess. Vastavalt vajadusele saab muuta dokumentide lisamise protsessi, defineerides uusi või eemaldades olemasolevaid samme. Dokumendi lisamise katkestamisel salvestatakse protsessi hetkeseis ning seda on võimalik hiljem samast kohast jätkata.

3.4 Fedora

Virginia ja Cornwelli Ülikooli poolt loodud vabavaraline tarkvara valmis aastal 2003 [FED]. Viimane uuendus avaldati jaanuaris 2007. Fedora kasutamine on reguleeritud ECL (Educational Community Licence) litsentsiga. Fedora on kodeeritud Javas, andmebaasiks sobivad McKoi, Oracle, PostgreSQL või MySQL.

Erinevalt DSpace'i ja Greenstone'i repositooriumist on Fedoral sisseehitatud dokumentide versioonide haldamine, mis annab ülevaate aja jooksul toimunud muudatustest. Samas puudub failide terviklikkuse kontroll.

Dokumentide metainfot saab kirjeldada erinevates formaatides. Vaikimisi

on selleks Fedora vajadustele kohandatud METS (Metadata Encoding and Transmission Standard).

Sarnaselt DSpace'i ja Greenstone'i repositooriumi süsteemidele toetab ka Fedora OAI-PMH standardit. Samas püsivate identifikaatorite realisatsiooni ei kasutata.

Kasutajate, rollide ja õiguste haldamine on paindlik. Rolle saab defineerida ja neid kasutajatele omistada. Kasutajate autentimist on võimalik teha ka üle LDAP (Lightweight Directory Access Protocol) protokolli. Lisaks saab ligipääsu reguleerida ka ip-aadressi põhiselt.

Üheks oluliseks miinuseks on täistekstiotsingu ja defineeritud töövooprotsessi puudumine. Otsida saab vaid repositooriumis olevate dokumentide metaandmetest, kasutades metamärke ja tõeväärtusloogikat.

3.5 Vastavus teadusgrupi vajadustele

Analüüsides võrdluse tulemusi, võib öelda, et funktsionaalsuse poolest on kõige tagasihoidlikum Fedora. Puudub materjali hulgi lisamise ja selle grupeerimise võimalus ning otsida saab metaandmetest, aga mitte täistekstist. Ühe olulise miinusena, mis käib kõigi kolme vaatluse all oleva süsteemi kohta, tuleb märkida kommenteerimise ja hindamise funktsionaalsuse puudumist.

Nii Greenstone kui ka DSpace võimaldavad andmebaasi sisu eksportida ja importida. Eksportimisel luuakse eraldi metainfo XML-failid, mida kasutatakse importimisel. Ainult failisüsteemist faile hulgi lisada ei saa.

Täistekstiotsingu tugi on olemas mõlemal viimatimainitul, kasutada saab nii metamärke kui ka tõeväärtusloogikat. Andmebaasi sisu grupeerimiseks on nii Greenstone'l kui ka DSpace'l *kogu* mõiste. Kogule saab alamkogusid luues anda hierarhilise struktuuri.

Arvestades ka teisi mitte otseselt teadusgrupi vajadustest tulenenud kriteeriume (Tabel 3.1), tuli sobivaimaks tunnistada DSpace. DSpace'i eelisteks Greenstone'i ees on paindlikum kasutajate ja õiguste haldus, sisu lehitsemise võimalus kasutajaliidesest ning defineeritud töövoog.

	Greenstone	DSpace	Fedora
Toetatavad andmebaasid	GDBM	PostgreSQL, Oracle	Mckoi SQL, PostgreSQL, Oracle, MySQL
Programmeerimiskeel	C++, Java, Perl	Java	Java
Mitmekeelsus	Jah	Jah	Ei
Metainfo standardid	Dublin Core, METS, MARC	Dublin Core	Dublin Core, METS, MARC
Standardite toetus	OAI-PMH	OAI-PMH, püsivad identifikaatorid	OAI-PMH,
Dokumentide import-eksport	Jah	Jah	Jah
Õiguste haldus	Fikseeritud 3 kasutajagruppi	Rollid, kasutajad, õigused	Ligipääsu reguleerimine ip-aadressi põhiselt
Sisu lehitsemine	Pole võimalik	Pealkirja, autori, kuupäeva järgi	Pole võimalik
Täistekstiotsing	Jah	Jah	Ei
Sisu algupärasuse säilitamine	Osaliselt	Jah	Osaliselt
Väliste autentimismeetodite tugi	Puudub	LDAP	LDAP
Töövoog	Puudub	Defineeritud töövoog	Puudub
Sisu grupeerimine	Jah	Jah	Ei
Kasutajaliides	Integreeritud veebirakendus	Integreeritud veebirakendus	Veebirakendus

Tabel 3.1: Digitaalsete dokumentide repositooriumite Greenstone'i, DSpace'i ja Fedora võrdlustabel

Peatükk 4

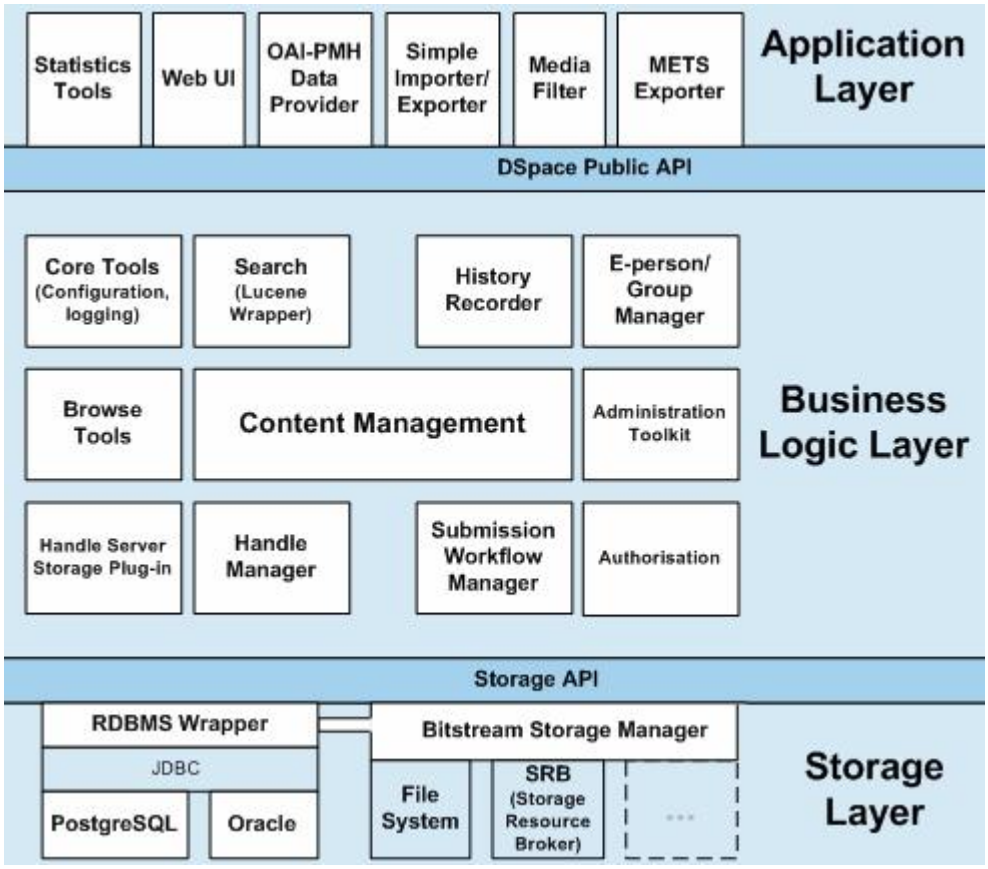
DSpace'i arhitektuuri ja funktsionaalsuse ülevaade

Kolme võrdluse all olnud digitaalse materjali andmebaasi tarkvarast sai valitud DSpace, kui sobivaim teadusgrupi BIIT vajadustele. Järgnevalt annan dokumentatsioonile tuginedes põhjalikuma ülevaate selle põhifunktsionaalsusest [DSD].

4.1 Arhitektuur ja andmemudel

Arhitektuuriliselt jaguneb DSpace kolmeks abstraktseks – rakenduse (ik *Application Layer*), äriloogika (ik *Business Logic Layer*) ja salvestuskihiks (ik *Storage Layer*) (Joonis 4.1). Rakenduse kihti kuuluvaks loetakse veebipõhine kasutajaliides, import-eksport skriptid ja statistika genereerimise vahendid. Sisu-, kasutajate- ning õiguste haldus, püsivate identifikaatorite ning ka otsingu- ja indekseerimisloogika paikneb äriloogika kihis. Andmebaasi ja failisüsteemiga suhtlemine on kapseldatud eraldi kihti ning see toimub läbi bitivoo salvestushalduri (ik *Bitstream Storage Manager*). Bitivoo salvestushalduri ülesanneteks on näiteks DSpace keskkonda üleslaaditud materjali talletamine failisüsteemis ning metainfo sisestamine andmebaasi.

Andmemudeli keskseks mõisteks on ühik (ik *item*, edaspidi kasutan sünonüümina mõistet dokument) (Joonis 4.2). Dokumendid kuuluvad *kogusse*, mis omakorda kuuluvad *kommuunidesse*. *Kogu* on mõeldud sarnase sisuga



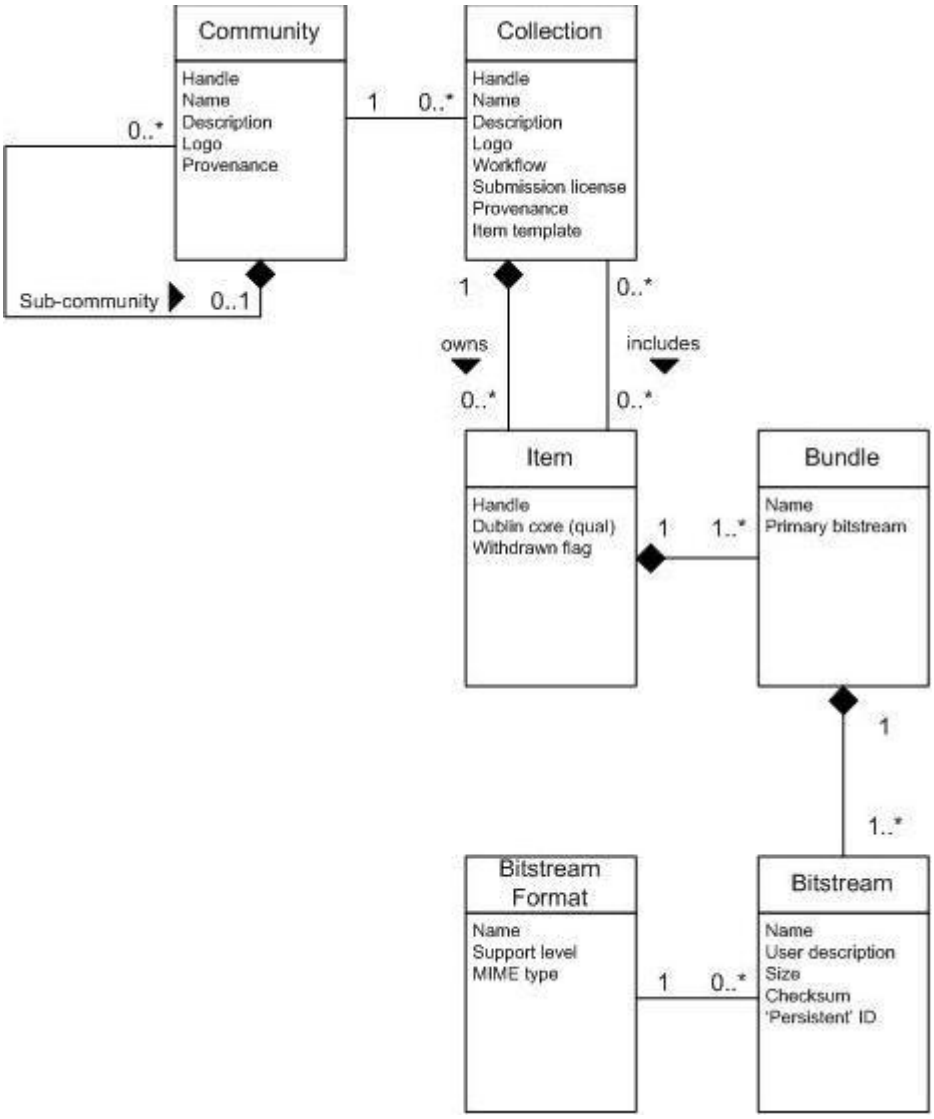
Joonis 4.1: DSpace kolmekihiline arhitektuur [DSD]

dokumentide grupeerimiseks. *Kommuun*, mille all võib mõista uurimisrühma, osakonda jne, on hierarhia kõrgeim tase.

Dokument koosneb *bitivoogude* (ik *bitstream*) hulgast, millesse kuulub vähemalt üks *bitivoog*. *Bitivoog* võib olla kas pildi- (jpg, gif, png), teksti- (pdf, doc), video- (avi, mpg) või mõnes muus formaadis fail.

4.2 Metainfo

Iga dokumendi kirjega on seotud metainfo, mille vaikimisi formaadiks on Dublin Core. Dublin Core standard määrab ära terminite hulga, mille väärtused konkreetset dokumenti kirjeldavad. Terminiteks on näiteks pealkiri, autor, avaldamise kuupäev, dokumendi päritolu määrav viide, bitivoo



Joonis 4.2: DSpace andmemudeli põhiseosed [DSD]

suurus jne. Metaandmeid hoitakse andmebaasis.

4.3 Kasutajate tuvastamine

Autentimine ehk kasutajatuvastus põhineb kasutajanime, milleks on e-posti aadress, ja parooli kombinatsiooni kontrollil. Alternatiivina on olemas ka LDAP ja X.509 sertifikaadiga autentimise tugi.

Vastavalt sellele, kas kasutaja on autenditud või mitte, on DSpace keskkonnas võimalik sooritada erinevaid tegevusi. Dokumendite vaatamiseks ja allalaadimiseks ei pea tingimata olema sisse loginud. Samas dokumentide lisamiseks, e-posti teavitajate aktiveerimiseks ja haldamistoimingute (õiguste, kasutajate lisamine ja muutmine) sooritamiseks tuleb kasutaja siiski tuvastada.

4.4 Õiguste süsteem

DSpace'i õiguste süsteem põhineb konkreetsete tegevuste (lugemine, lisamine, muutmine, kustutamine, administreerimine) seostamisel objektide ja kasutajatega (Tabel 4.1). Administreerimise lihtsustamiseks on olemas rollid, mis sisaldavad endas õiguste komplekte ning mida saab kasutajatele määrata. Algselt on defineeritud kaks rolli – administraatorid ja anonüümsed kasutajad.

Õigusi saab anda nii üksikule kasutajale kui ka rollile. Õigused ei laiene alamobjektidele ja seetõttu tuleb neid anda ilmutatult. Näiteks kui kasutajal on dokumendi lugemisõigus, kuid tal puudub selle dokumendi bitivoo lugemisõigus, siis dokumendi allalaadimine ebaõnnestub.

4.5 Dokumentide lisamine

Dokumentide lisamine on mitmesammuline protsess (Joonis 4.3), mida saab algatada veebipõhisest kasutajaliidesest või käsurealt skriptiga.

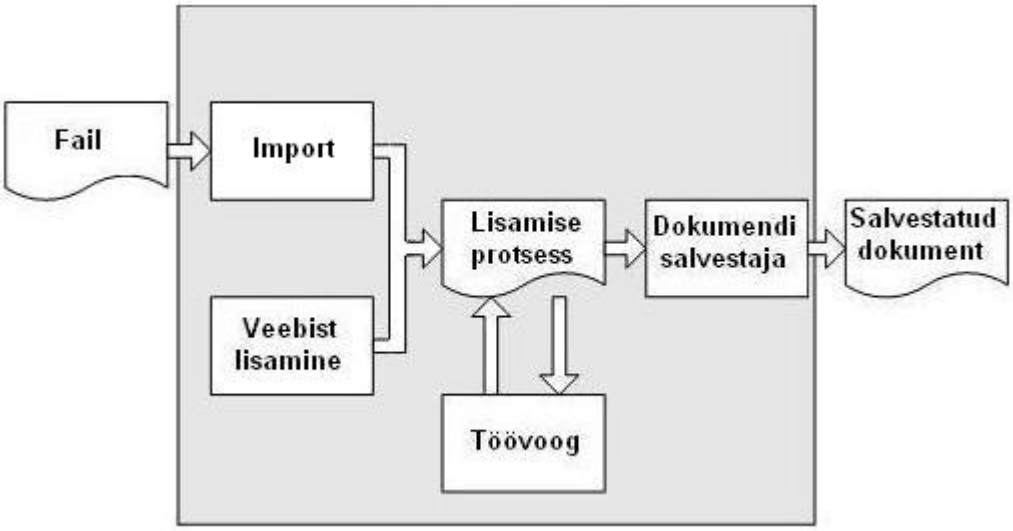
Käsurealt skriptiga algatatav protsess on põhiliselt mõeldud DSpace repositooriumite vaheliseks dokumentide ekspordiks-impordiks. Import eeldab

Objekt	READ	WRITE	ADD	REMOVE	ADMIN
Kommuun			+	+	
Kogu			+	+	+
Dokument	+	+	+	+	
Bitivoogude hulk			+	+	
Bitivoog	+	+			

Tabel 4.1: DSpace õigused ja objektid. READ - vaatamisõigus, WRITE - muutmisõigus, ADD - lisamisõigus, REMOVE - kustutamisõigus, ADMIN - administreerimise õigus

XML-kujul metainfo faili olemasolu, mis tekitatakse dokumendi ekspordil. Dokumentide eksport-import funktsionaalsust saab kasutada näiteks varundamiseks.

Kasutajaliidesest dokumentide lisamine sisaldab endas ka mitmesammulist töövoogu. Järjestikuste sammude jooksul sisestatakse metainfo, laaditakse üles fail, kontrollitakse sisestatut ning lõpuks salvestatakse. Protsessi on võimalik igal sammul katkestada ning hiljem jätkata. Kõik pooleli jäänud dokumentide lisamised on näha veebiliideses pärast kasutaja sisselogimist.



Joonis 4.3: Dokumentide lisamise protsess [DSD]

Igale *kogule* on võimalik määrata üks või mitu ülevaatajat (ik *supervisor*). Ülevaataja ülesandeks on lisatud materjali enne repositooriumisse salvestamist kontrollida ja juhul, kui kõik on korrektne, ka kinnitada. Tagasilükkamise korral peab kasutaja tegema täiendusi ja kinnitamiseks uuesti esitama.

4.6 Identifikaatorid

Üks levinumaid viise internetis olevale ressursile, olgu selleks siis veebileht, pilt, fail vms, viitamiseks on URL (Universal Resource Locator). Selle üheks puuduseks on jäik seos geograafilise asukohaga. Ressursi tõstmisel ühest serverist teise, pole see enam vanalt aadressilt leitav.

Püsivate viitade tagamiseks kasutatakse CRNI (Corporation for National Research Initiatives) identifikaatorite süsteemi, mis koosneb ühest globaalsest registrist (GHS – Global Handle Registry) ning paljudest kohalikest teenustest (LHS – Local Handle Service) [CRN]. Globaalses registris hoitakse viiteid lokaalsetele teenustele, milles omakorda säilitatakse ning hallatakse püsivate identifikaatorite ja URL-ide vastavusi (Joonis 4.4).

Püsiv identifikaator koosneb globaalsest prefiksist ning lokaalsest sufiksist. Eesliite abil tuvastatakse teenus, mis identifikaatorile vastava URL-i tagastab (Joonis 4.5). Tasub märkida, et antud süsteem võimaldab ühe identifikaatoriga viidata ühe ressursi erinevatele formaatidele (PDF, XML jne) (Joonis 4.5).

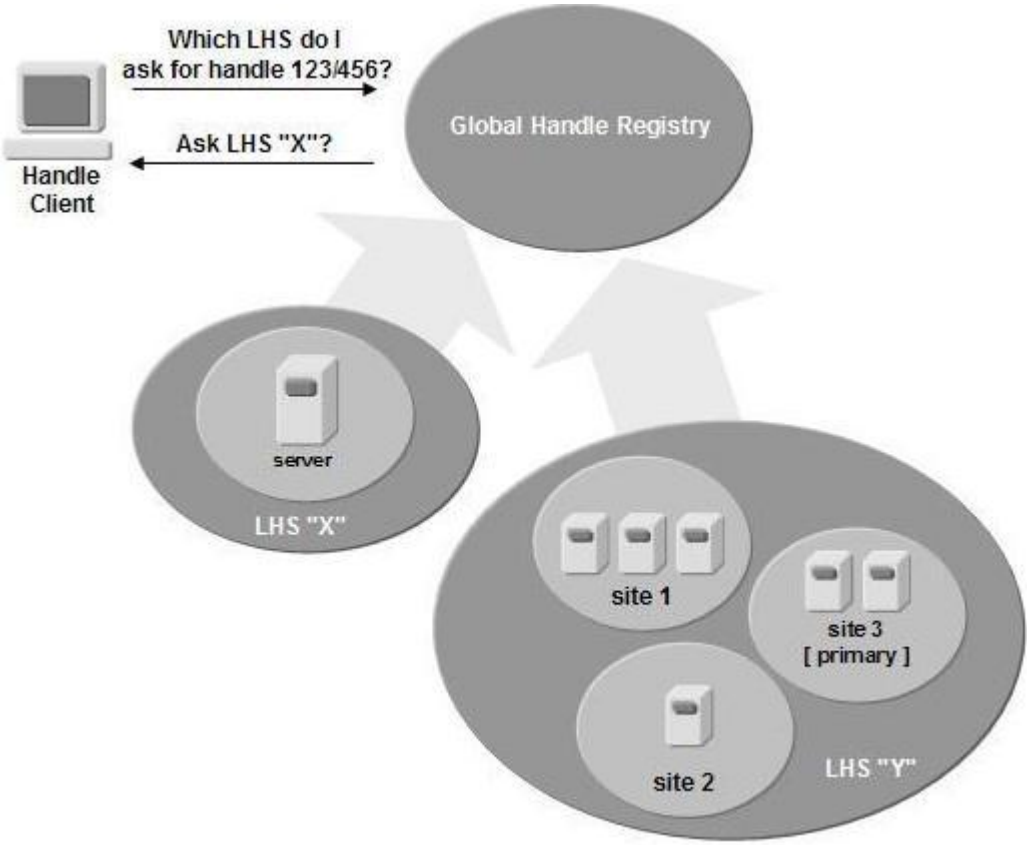
Püsiva identifikaatori täiskujusid on kaks:

- hdl:10.12493/12394
- http://hdl.handle.net/10.12493/12394

Esimene neist on kasutatav veebibrauseris, mis toetab HDL protokolli. Kuna enamik veebilehitsejaid seda ei toeta, siis on võimalik kasutada HTTP protokollil põhinevat täiskuju.

4.7 Dokumentide sirvimine ja otsing

Dokumente saab sirvida pealkirja, lisamise kuupäeva ja autori nime järgi repositooriumi, *kommuuni* või ka *kogu* tasemel. Selle saavutamiseks haldab



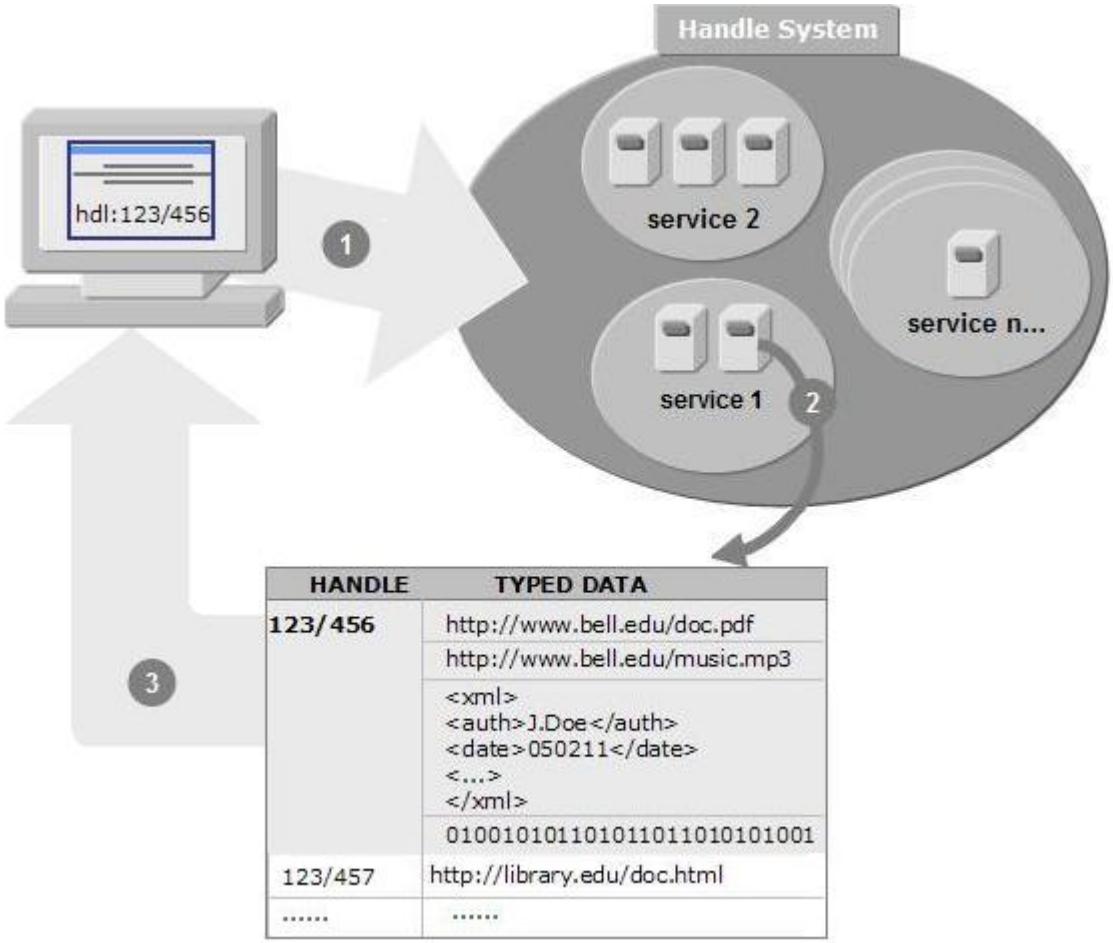
Joonis 4.4: CRNI identifikaatorite süsteemi kahetasemelisus. Globaalne register annab päringu vastusena lokaalse teenuse, mis omakorda tagastab küsitud dokumendi [CRN]

DSpace eraldi indekseid.

Otsida on võimalik nii metainfo väljade väärtuste järgi kui ka dokumendi sisust. Sarnaselt dokumentide sirvimisele saab ära määrata, kas otsitakse tervest repositooriumist või mõnest *kommuunist*, *kogust*.

Indekseerimiseks ja päringute sooritamiseks kasutatakse Javas realiseeritud Lucene [LUC] otsimootorit. Kasutada saab metamärke "?" ja "*". Esimene metamärk tähendab 1, teine 0 või enama sümboli asendust (Tabel 4.2, näited 1, 2, 3).

Sarnaste sõnade otsimist saab teostada hägusa otsinguga (ik *fuzzy search*), sümbolit "~" kasutades. (Tabel 4.2, näide 4). Sarnasuse määramisel kasutatakse sõnede teisenduskaugust (Levenšteini kaugus). Sarnasuse saab ise mää-



Joonis 4.5: CRNI identifikaatorite lahendumine [CRN]

rata vahemikus 0 kuni 1 (Tabel 4.2, näide 5), vaikimisi väärtuseks on 0,5.

Ligiduse otsing (ik *proximity search*) tähendab sõnade otsimist, mis on üksteisest kindlal kaugusel. Selleks tuleb fraasi (kahe või enama sõna) järel kasutada sümbolit "~" ja järjestikuste sõnade arvu, mille hulka otsitavas fraasis olevad sõnad peavad tekstis kuuluma (Tabel 4.2, näide 6).

Vahemikotsingu (ik *range search*) teostamiseks tuleb defineerida kaks väärtust, mille vahele jäävatele väärtustele vastavad dokumendid tagastatakse. Alg- ja lõppväärtused võivad olla kas kaasa- või välja arvatud (Tabel 4.2, vastavalt näide 7 ja 8).

Otsinguterminite kombineerimiseks toetatakse tõeväärtusoperaatoreid "AND" (konjunktsioon), "+" (termini olemasolu nõudmine tekstis), "OR" (disjunkt-

Nr	Otsingutermin	Otsingule vastav tekst
1.	te?t	test, text
2.	test*	test, tester
3.	tes*r	tester
4.	test~	text, best
5.	test~0.8	best
6.	"red blue"~4	red car and blue van
7.	title:[a TO c]	a, b, c
8.	title:{a TO c}	b
9.	red AND (blue OR green)	red car, green van
10.		red car, blue van

Tabel 4.2: Otsingu võimalused

sioon), "-" (termini olemasolu keelamine tekstis) ja "NOT" (eitus). Grupeerimiseks tuleb kasutada sulge (Tabel 4.2, näide 9 ja 10).

4.8 Metainfo jagamine

OAI-PMH on Open Archive Initative poolt loodud metainfo jagamise ja kogumise protokoll [OAI]. Standard näeb ette kahte osapoolt – teenusepakkujat ja teenuse tarbijat. Teenuse tarbija teeb repositooriumisse parametriseeritud HTTP päringu, mille peale tagastatakse metainfot sisaldav XML.

DSpace toetab OAI-PMH metainfo kogumise protokolli olles teenusepakkuja rollis ning võimaldades välistel rakendustel pärida repositooriumis säilitatavate dokumentide metaandmeid.

4.9 Automaatne teavitus

DSpace võimaldab kasutajal registreerida *kogudele* teavitajaid, et olla kursis lisandunud ja muudetud dokumentide kohta. Uue dokumendi lisandumisel või olemasoleva muutmisel saadetakse süsteemi poolt kasutajale e-kiri.

4.10 Statistika

Statistika genereerimiseks kasutatakse DSpace'i süsteemilogisid, aga ka andmebaasi. Andmebaasis säilitatakse näiteks dokumentide allalaadimised kasutajate poolt. Kogu selle info põhjal saab genereerida järgmisi statistilisi aruandeid:

- *kommuunide*, *kogude*, dokumentide vaatamiste ja allalaadimiste arv
- väliste süsteemide poolt sooritatud OAI-PMH päringute arv
- kasutajate sisselogimised
- populaarseimad otsingud

Minimaalseks perioodiks on kuu ja maksimaalseks ajavahemik alates repositooriumi loomise hetkest. Lisaks saab genereerida aruandeid repositooriumis säilitatavate dokumentide arvu kohta *kogude* ja *kommuunide* lõikes.

Peatükk 5

Lisatud funktsionaalsus

Käesoleva töö kõige mahukam osa oli DSpace keskkonna teadusgrupi BIIT vajadustele sobivaks muutmine ning selle kasutusele võtmine. Selleks tuli realiseerida hulk funktsionaalsust, mis senises süsteemis puudus. Kogu kirjutatud funktsionaalsus sai vormistatud programmikoodi "paikadena" (ik *patch*) ja seda saab paigaldada uutele ametlikult väljalastud DSpace'i versioonidele. Sellise töömeetodi üheks puuduseks on asjaolu, et programmikoodi paigad tuleb pärast iga uue DSpace'i versiooni valmimist üle vaadata, testida ja vigade ilmemisel parandada.

5.1 Failide hulgi lisamine

Asjaolu, et aja jooksul oli teadusgrupi liikmete arvutitesse kogunenud hulk loetud ja ka produtseeritud kirjandust PDF-failidena, tingis vajaduse neid faile hulgi lisada. Seejuures oli lihtne ja mugav seda teha just käsurealt.

Hulgi lisamise nõude rahuldamiseks realiseerisin shelli skripti *batch-import.sh* ning Java klassi *BatchImport.java*, milles on kogu importimise funktsionaalsus. Skripti sisendparameetriteks on:

- -c <*kogu identifikaator*> – *kogu*, kuhu dokument lisatakse
- -e <*kasutajanimi*> – DSpace kasutajanimi
- <*fail*> – lisatav fail või failid

Skripti käivitamiseks tuleb Unixi käsurealt sisestada:

```
sh batch-import -c Test -e teints@math.ut.ee paper.pdf
sh batch-import -c Test -e teints@math.ut.ee *.pdf
```

Siinkohal tasub märkida, et lisatavad failid võib ette anda shelli regulaaravaldisena. Näiteks avaldis **.pdf* impordib repositooriumisse kõik jooksvas kataloogis olevad pdf-laiendiga failid.

Failide hulgi lisamise üheks miinuseks on metainfo puudulikkus. Faili üleslaadimisel salvestatakse metaandmetesse vaid selle nimi, suurus ja loomise kuupäev. Artikli sisu kohta käiv metainfo tuleb kasutajaliidesest käsitsi lisada, mis on aeganõudev ja seetõttu tülikas.

5.2 Metainfo automaatne kogumine

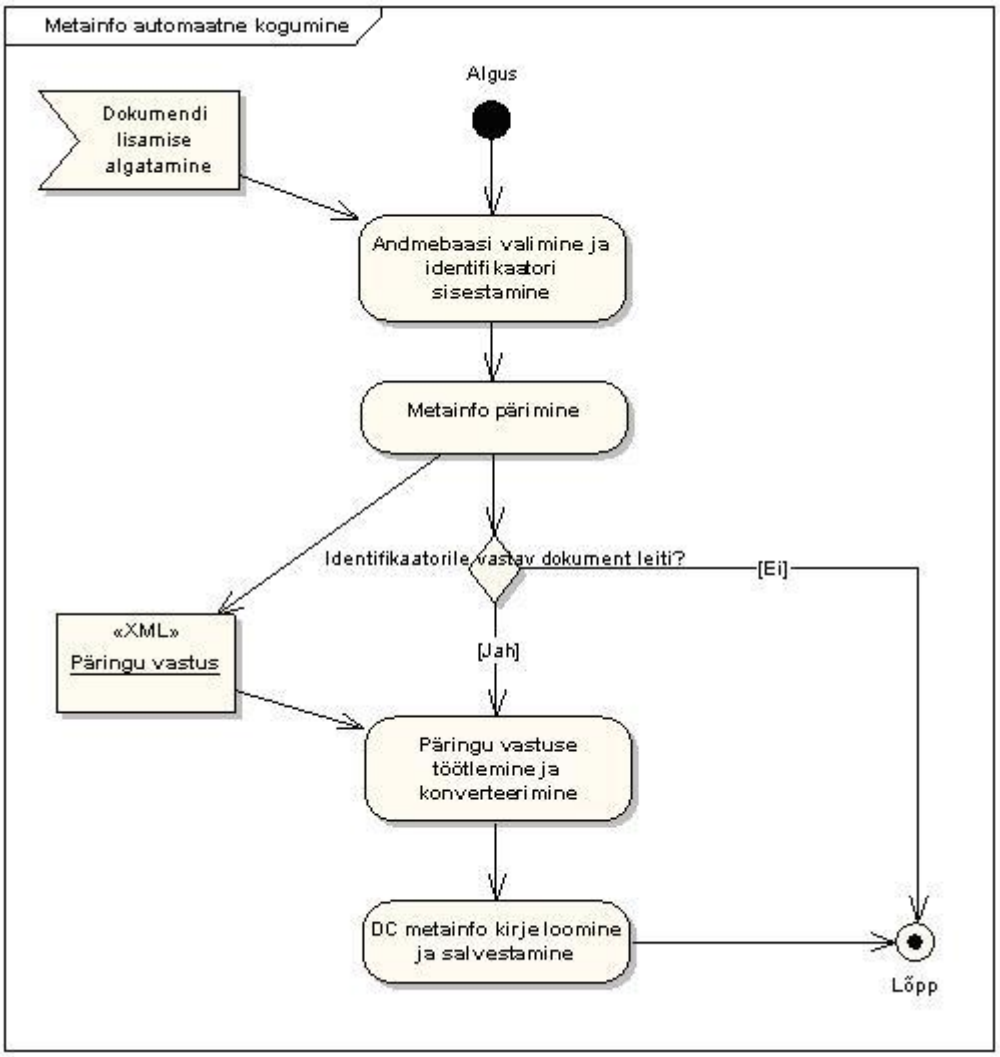
Teadusgrupi BIIT poolt kasutatava kirjanduse päritolu analüüsides selgus, et enamik artikleid pärineb suurtest teaduskirjanduse andmebaasidest nagu Pubmed Central, Biomed Central, Citeseer või Pubmed. Neist kolme esimese võimalusi uurides leidsin, et nad toetavad OAI-PMH protokolli ja seega saaks metainfot programselt pärida ning BIIT-i DSpace repositooriumisse salvestada. Sellest tulenevalt tekkis mõte metainfo automaatse kogumise ja DSpace repositooriumisse salvestamise realiseerimiseks (Joonis 5.1). Kasutaja peab teadma vastavas andmebaasis oleva dokumendi identifikaatorit, mille metainfot koguda soovitakse.

Metainfo kogumiseks kasutasin OAI-PMH standardis spetsifitseeritud päringut *GetRecord*, mis on ette nähtud kindla artikli metaandmete pärimiseks:

```
http://www.pubmedcentral.nih.gov/oai/oai.cgi?verb=GetRecord&
identifier=oai:pubmedcentral.nih.gov:13900&metadataPrefix=oai_dc
```

Tegemist on HTTP GET tüüpi päringuga, mille parameetriteks on:

- verb – päringu tüüp
- identifier – dokumendi identifikaator
- metadataPrefix – tagastatava metainfo formaat



Joonis 5.1: Dokumendi lisamine metainfo automaatse kogumisega

Päringu tulemusel tagastatakse dokumendi metainfo XML kujul (Lisa 1).

Pubmedil OAI-PMH tugi puudus. Selle asemel oli olemas veebiteenus *efetch*, mis kasutamise ja eesmärgi poolest on analoogne OAI-PMH *GetRecord* päringule:

```
http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed
&id=1174891&retmode=xml
```

Päringu parameetrid on:

- db – andmebaas
- id – dokumendi identifikaator
- retmode – vastuse formaat; võimalikud variandid: xml, html, text ja asn.1

Päringu tulemusel tagastatav metainfo XML (töö lisas 3) pole aga Dublin Core formaadis ning seetõttu tuli defineerida reeglid, mille alusel metaandmed DSpace'i repositooriumile sobivaks konverteerida.

5.3 Artiklite lisamise protsesside defineerimine ja seadistamine

Üheks DSpace oluliseks eeliseks teiste digitaalse materjali andmebaaside tarkvara ees oli täpselt ette määratud töövoog. See koosneb metainfo käsitsi sisestamise, faili üleslaadimise, andmete korrektsuse kontrolli ning litsentsi määramise sammudest. Sammude arv ja nende järjekord on jäigalt määratud ning seadistamine ilma programmikoodi muutmata võimatu.

Pärast metainfo automaatse kogumise ja salvestamise funktsionaalsuse realiseerimist tekkis vajadus see ühtsesse töövoogu integreerida. Samas pidi kasutajale jääma võimalus valida, kuidas ta metainfot lisab. Sisuliselt oli vaja võimalust defineerida *kogule* kaks erinevalt kulgevat tegevuste jada.

Algselt ei võimaldanud DSpace defineerida kahte või enamat töövooprotsessi *kogu* kohta. DSpace wiki keskkonda lugedes selgus, et osa vajaminevast funktsionaalsusest on realiseerinud Tim Donohue, üks DSpace aktiivsetest arendajatest [SUB]. Tema poolt loodud programmikoodi täienduste tulemusel sai määrata töövooprotsessi samme ning nende järjekorda. Samas polnud siiski võimalik defineerida *kogule* rohkem kui ühte töövoogu.

Võtsin kasutusele Tim Donohue programmikoodi ning pärast mõningaid täiendusi sai *kogule* omistada mitut erinevat dokumendi lisamise tegevuste jada. Kogu seadistus on kirjeldatud XML-failis (Lisa 4). Esimeses blokis (*submission-map*) määratakse ära *kogud* ja nende töövood. Teises blokis (*step-definitions*) on defineeritud töövoo sammud ning kolmandas blokis (*submission-definitions*) kirjeldatakse ära töövood.

5.4 Dokumentide tõstmine ühest *kogust* teise

Üheks teadusgrupi BIIT vajaduseks oli, et repositooriumisse salvestatud dokumente peab saama klassifitseerida ja grupeerida. Selleks on mõeldud *kommuunid* ja *kogud*, millele saab defineerida *alamkommuune* ja *-kogusid*.

Failide hulgi lisamisel talletatakse materjali repositooriumisse minimaalse metainfoga eeldusel, et dokumenti hiljem kas käsitsi või metainfo automaatse kogumise funktsionaalsust kasutades täiendatakse. Mõistlik oleks selliseid dokumente hoida eraldi "*töökogus*" ja need alles pärast andmete korrigeerimist sealt ära tõsta. DSpace kasutamisel võib ka juhtuda, et mõni dokument salvestatakse ebasobiva *kogu* alla ja on vaja seetõttu ümber tõsta.

Dokumentide tõstmine kasutajaliideses ühes *kogust* teise polnud DSpace'is algselt kahjuks võimalik ning tuli seega realiseerida. Esialgu piisas ükshaaval ümbertõstmisest, mida saab nüüd teha dokumendi andmete muutmise ekraanilt. Selleks tuleb kasutajal vaid määrata *kogu*, kuhu dokument tõsta (Joonis 5.2). Andmebaasis hoitakse igal dokumendil viita *kogule*, kuhu ta kuulub. Ümbertõstmisel muudetakse lihtsalt viida väärtust.

5.5 PDF failide vaatamine tekstina

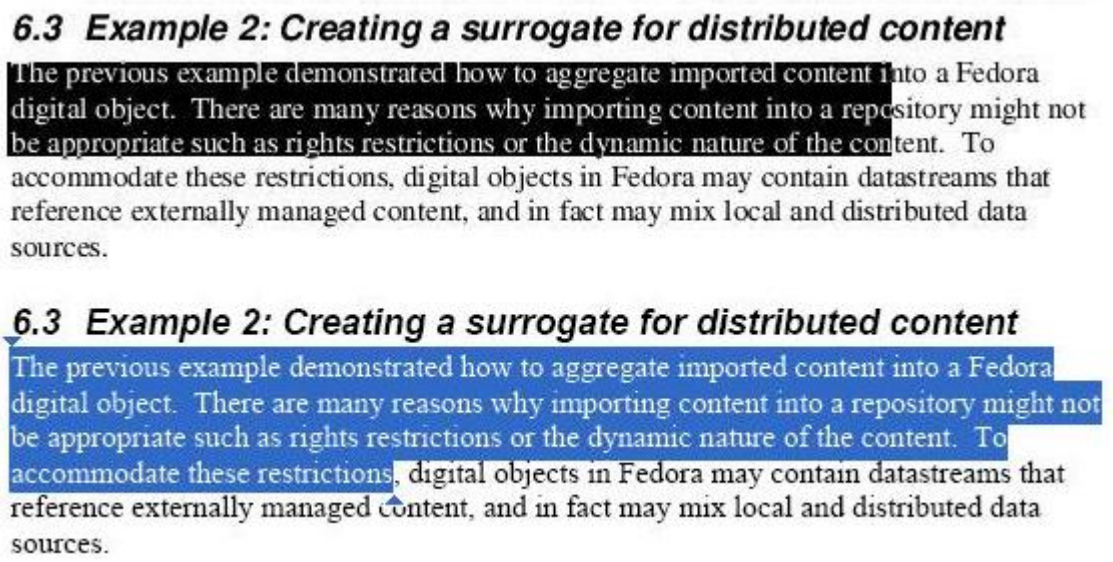
PDF-formaadis failide avamine veebilehitsejas võtab kauem aega kui sama infot sisaldava tekstifaili avamine. Erinevuse põhjuseks on PDF-faile lugeda suutva programmi esmakordse käivitamise aeglus.

Samuti võib probleemiks osutuda teksti kopeerimine. Näiteks Unixi-laadsete operatsioonisüsteemide PDF klientprogrammiga *xpdf* saab märgistada vaid neljakandilist blokki, samas kui *Adobe Acrobat Reader* oskab ära märkida ka pool rida.

Üheks võimaluseks kirjeldatud ebamugavusi vältida oli PDF-failist teksti eraldamine ja selle kuvamine veebilehitsejas. Teksti avamine veebibrauseris on oluliselt kiirem kui PDF faili korral ning märgistada saab vajadusel ka ainult poolteist rida.

PDF-failidest teksti eraldamiseks on olemas mitmeid konverteerimisprogramme. DSpace kontekstis sobis selleks Javas realiseeritud vabavaraline PDFBox [PDFB], mis mõeldud PDF-formaadis failide töötlemiseks. DSpace ka-

Joonis 5.2: Dokumendi tõstmine ühest *kogust* teise



Joonis 5.3: PDF klientprogrammidega (üleval *xpdf*, all *Adobe Acrobat Reader*) teksti märgistamine

sutajaliideses dokumendi detailvaatesse sai lisatud link "View As Text", mis kuvatakse vaid pdf-laiendiga failidel. Lingile vajutades avatakse uus veebibrauseri aken, kus kuvatakse faili sisu teksti kujul.

5.6 Kommenteerimine ja soovitamine

Soovitamine seisneb kasutajale huvipakkuvate objektide tuvastamises. Üks laiemalt levinud meetod põhineb kasutajate käitumise ajaloo analüüsil [DK04]. Eristatakse kasutaja- ja mudelipõhist lähenemist.

Kasutajapõhine lähenemine käsitleb iga kasutajat kui sarnaselt käituvate indiviidide grupi liiget. Grupi liikmete huvide põhjal saab igale grupi liikmele teha soovitusi. Mudelipõhisel lähenemisel analüüsitakse kasutajate käitumisharjumisi ning leitakse objektide vahelised seosed ning nende tingimuslikud tõenäosused. Objektide o_1 ja o_2 korral püütakse tuvastada, kui suur on tõenäosus, et kasutajat huvitab objekt o_2 eeldusel, et teda huvitas ka objekt o_1.

Kommenteerimise ja soovitamise funktsionaalsus DSpace rakendusele oli

juba "koodipaigana" realiseeritud [MIN]. Käesoleva töö raames sai see mõningate täiendustega integreeritud teadusgrupi BIIT DSpace rakendusse. Kasutajatele soovituste andmiseks on kasutatud kahte algoritmi: "Sotsiaalse informatsiooni filtreerimine" (ik *Social Information Filtering* ja "Item-Based Top-N" algoritmi.

Nii "Sotsiaalse informatsiooni filtreerimine" kui ka "Item-Based Top-N" algoritm püüavad leida ja soovitada konkreetsele kasutajale artikleid. Soovitamise aluseks on artiklite allalaadimiste info põhjal arvutatud olulisuse skoor. Soovitamise funktsionaalsuse täpsust on keeruline hinnata põhjusel, et nad eeldavad piisava hulga andmete (allalaadimiste arvu) olemasolu. Töö kirjutamise hetkel oli andmeid analüüsimiseks veel liiga vähe.

5.6.1 Sotsiaalse informatsiooni filtreerimine

Sotsiaalse informatsiooni filtreerimine (ik *Social Information Filtering*) on oma olemuselt kasutajapõhise lähenemisega kasutaja käitumisharjumuste informatsiooni analüüsimine. Algoritm koosneb kahest sammust. Esimeses tuvastatakse sarnaste huvidega kasutajad ning teises leitakse soovitused nendele kasutajatele huvipakkunud objektide hulgast.

DSpace kontekstis on objektideks dokumendid ja kasutaja huvi mõõdikuks dokumendi juurde kuuluvate failide allalaadimine. Tähistame kasutaja i poolt allalaaditud failide hulga F_i, kus $i = 1, \ldots, m$. Kasutajate i ja j huvide sarnasuse skoor $Sim(F_i, F_j)$ arvutatakse valemiga:

$$Sim(F_i, F_j) = \frac{2 \cdot \mid F_i \cap F_j \mid}{\mid F_i \mid + \mid F_j \mid}$$

Seega on kahe kasutaja i ja j huvid täiesti sarnased ($Sim(F_i, F_j) = 1$) siis, kui mõlemad on alla laadinud täpselt samad failid ning täielikult erinevad ($Sim(F_i, F_j) = 0$), kui ei leidu ühtegi faili, mida oleks alla laadinud nii kasutaja i kui ka j.

Järgmisena valitakse n, $n \leqslant m$ vaatluse all olevale kasutajale sarnaseimate huvidega kasutajat. Leitakse dokumendid, mida see kasutaja pole alla laadinud ning soovitataksegi neid.

5.6.2 Item-Based Top-N algoritm

Antud algoritm kasutab mudelipõhist lähenemist, mille korral defineeritakse sarnasuse mõõduna tingimuslik tõenäosus objektide vahel [DK04]. Antud töö kontekstis on objektideks dokumendid ning mõõdikuks dokumentide allalaadimiste sagedus.

Olgu meil dokument d_i, $i = 1, \ldots, m$, millele tuleb leida sarnaseid dokumente. Sarnasuse mõõduks on tõenäosus, et kasutaja laadib alla dokumendi d_j faile eeldusel, et ta on laadinud alla dokumendi d_i faile (Valem 5.1).

$$sim(d_i, d_j) = P(d_j \mid d_i) = \frac{Freq(d_i d_j)}{Freq(d_i)} \tag{5.1}$$

Siinkohal tuleb ära märkida, et tegemist on asümmeetrilise algoritmiga. Olgu meil kaks dokumenti d_i ja d_j, millest esimest on alla laetud märkimisväärselt rohkem kordi kui teist. Võttes eelduse aluseks dokumendi d_i, saame d_i ja d_j sarnasuseks oluliselt väiksema suuruse, kui eelduse aluseks oleks dokument d_j.

$$sim(d_i, d_j) = P(d_j \mid d_i) \neq P(d_i \mid d_j) = sim(d_j, d_i) \tag{5.2}$$

Asümmeetrilisusest tuleneva ebavõrdust saab vähendada normaliseerimisfaktorit α kasutades. Sellisel juhul on sarnasuse leidmiseks kasutatav valem järgmine:

$$sim(d_i, d_j) = \frac{Freq(d_i d_j)}{Freq(d_i) \cdot (Freq(d_j))^{\alpha}} \tag{5.3}$$

Tasub märkida, et kui $\alpha = 0$, siis on meil tegemist valemiga (Valem 5.1) ja kui $\alpha = 1$, saame sümmeetrilise algoritmi, kus

$$sim(d_i, d_j) = P(d_j \mid d_i) = P(d_i \mid d_j) = sim(d_j, d_i) \tag{5.4}$$

5.7 Artiklite hindamine

Võimaldamaks kasutajatele hinnangute andmist andmebaasis oleva materjali kohta, realiseerisin hindamise funktsionaalsuse. Hinnangu skaala (alg-

Joonis 5.4: Keskmise hinde kasutamine otsingu tulemuste sorteerimiseks

selt 5-palli süsteemis) on defineeritud andmebaasis ja seda saab vastavalt vajadusele lihtsalt muuta.

Keskmine hinne S leitakse geomeetrilise keskmise arvutamisel:

$$S = \sqrt[n]{q_1 \cdot q_2 \cdot \ldots \cdot q_n} \tag{5.5}$$

Artiklitele antud hinnangute pealt arvutatud keskmist hinnet kasutatakse otsingutes tulemuste sorteerimiseks (Joonis 5.4).

5.8 Indeksi optimeerimine

Lisamiste ja kustutamiste tulemusena aja jooksul indeksi suurus kasvab ning otsimine muutub aeglasemaks [LUC]. Repositooriumist kustutatud dokumendid eemaldatakse indeksist alles optimeerimisel.

Indeksi optimeerimiseks kasutasin DSpace'i olemasolevat shelli skripti *index-all.sh*. Skripti tööd testides selgus, et teatud juhtudel optimeerimine ebaõnnestub. Probleemi uurides osutus vea põhjuseks indeksi segmentide kustutamine. Kui mõni DSpace'i kasutaja parajasti andmebaasist midagi otsis ja indeksi segmendist andmeid loeti, ebaõnnestus selle kustutamine. Lahenduseks oli DSpace rakenduse töö peatamine optimeerimise ajaks ning pärastine taaskäivitamine.

Indeksi optimeerimist teostatakse öösiti kell 4. Unixi-laadsetes keskkondades on perioodiliselt käivitatavate tegevuste defineerimiseks *crontab* käsk (Näide 5.8). Koostasin skripti *index-dspace-content.sh*, mis peatab rakenduse töö, optimeerib indeksi ning seejärel rakenduse taaskäivitab. Skripti töötamise ajal standardväljunditesse *stdout* ja *stderr* kirjutatav info logitakse faili *index-dspace-content.out*.

Näide 5.8: crontab käsk indeksi optimeerimiseks kell 4 öösel

```
00 4 * * * /group/software/teino/crontab/index-dspace-content
1>>/group/software/teino/crontab/logs/index-dspace-content.out
2>\&1
```

Peatükk 6

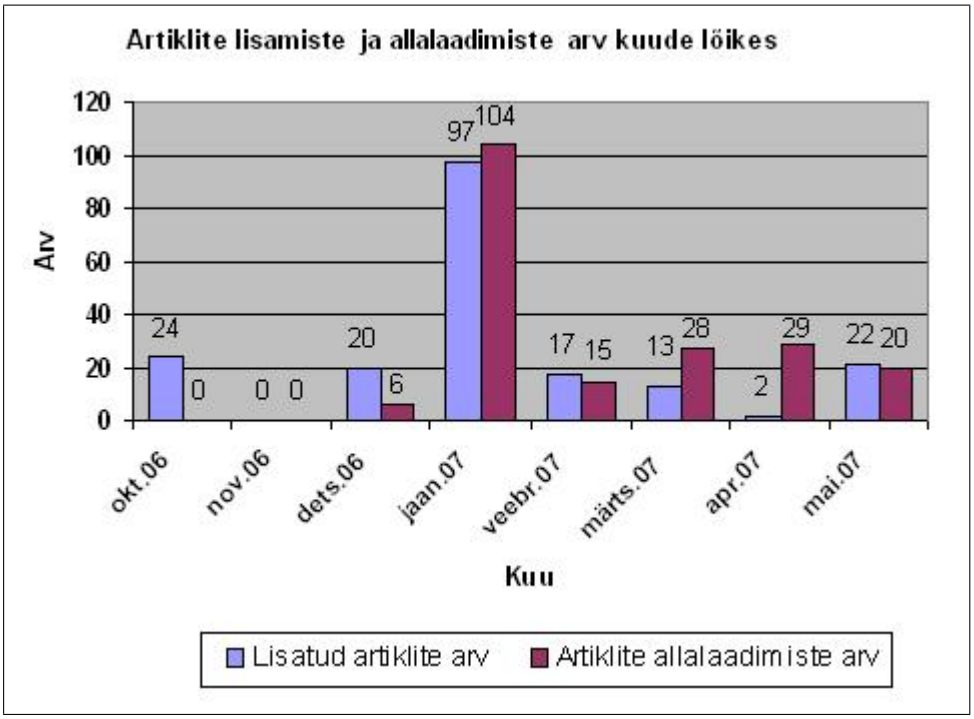
DSpace'i statistika ja kasutusaktiivsus uurimisrühmas BIIT

Töö kirjutamise hetkeks oli end DSpace kasutajaks registreerinud enamik uurimisrühma BIIT liikmetest so 22 inimest 30-st. Andmebaasi oli defineeritud 28 *kogu*, millesse lisatud 203 faili. Praktiliselt kõik failid on PDF-formaadis (Tabel 6.1). Seda eeldasin teadusgrupi vajaduste analüüsi põhjal ning sain statistikale tuginedes kinnituse. Alates 2006. aasta detsembrist oli andmebaasis salvestatud kirjandust alla laetud 195 korda 14 erineva kasutaja poolt.

Faili tüüp	Failide arv
PDF	198
Microsoft Powerpoint	3
JPEG	2

Tabel 6.1: Failide arv tüübi järgi

Kommenteerimise ja hindamise funktsionaalsuse kasutusaktiivsust uurides tuleb kahjuks tõdeda, et see pole peaaegu üldse rakendust leidnud. Kommenteeritud on vaid kahte dokumenti ühe kasutaja poolt ning hinnanguid on andnud kolm erinevat kasutajat neljale artiklile.



Joonis 6.1: Artiklite lisamiste ja allalaadimiste arv DSpace'i kuude lõikes

Vaadates artiklite lisamiste ja allalaadimiste arvu kuude lõikes (Joonis 6.1), võib öelda, et kirjanduse lisamiste arv on kuude lõikes üsna ebastabiilne. Kindlasti tuleb veel teha uurimisgrupi siseselt aeg-ajalt selgitustööd ja veenda selle liikmeid aktiivsemalt DSpace'i kasutama. Allalaadimiste vähesus on seletatav kirjanduse suhteliselt väikese hulgaga. Tõenäoliselt aja jooksul suurenev artiklite arv mõjutab seda näitajat positiivses suunas.

Kokkuvõte

Publitseeritava teaduskirjanduse säilitamiseks ja levitamiseks on mitmeid andmebaase, mida uurimisrühmad oma töös kasutavad. Samas on ka materjale, mida on vaja talletada teadusgrupi siseselt. Käesoleva töö eesmärgiks oli analüüsida uurimisgrupi BIIT vajadusi ning luua teadusgrupi sisese kirjanduse säilitamiseks ja levitamiseks kasutatav keskkond.

Kolme vabavaralise digitaalse materjali repositooriumi tarkvara võrdluse tulemusel selgus, et olemasoleva funktsionaalsuse põhjal vastab teadusgrupi BIIT vajadustele kõige paremini DSpace.

Lähtudes uurimisgrupi vajadustest, sai realiseeritud hulk lisafunktsionaalsust. Tehtud täienduste tulemusel saab faile hulgi lisada ja metaandmeid suurtest teaduskirjanduse andmebaasidest automaatselt koguda ning DSpace keskkonda salvestada. Oluliste ja heade artiklite eristamiseks sai integreeritud kommenteerimise ja soovitamise funktsionaalsus ning lisatud hindamine.

Kasutusaktiivsust analüüsides tuli tõdeda, et DSpace'i juurutamine pole läinud just kõige ladusamalt ning selle nimel tuleb ennekõike teavitustööd tehes veel vaeva näha. Kahtlemata tuleb funktsionaalsust edasi arendada, et veelgi lisada praktilist väärtust. Ühe näitena võiks tuua metainfo põhjal BibTex'i kirjete genereerimise. Samuti oleks huvitav püüda tuvastada olulisi artikleid võttes arvesse erinevaid kriteeriume.

Literature database

Bachelor Thesis

Marten Teino

Abstract

Huge amount of digital information in science including articles, papers, books is being produced daily. This information has to be stored while keeping in mind the fact that it has to be accessible for being useful. There are public databases like Pubmed, Citeseer, ACM for this purpose. Unfortunately these public databases are for published material only. Reports, presentations, paper drafts should be stored as well.

The Bioinformatics, Algorithmics, and Data Mining research group (BIIT) is operating at the University of Tartu. So far there has been no environment to store and distribute read or produced papers and articles. The main goal of this bachelor thesis was to find out the needs of the research group, implement and deploy a literature database.

Greenstone, Fedora and DSpace are three open-source digital repository systems to capture, store, index, preserve and distribute digital research material. Based on the comparison, the best candidate for BIIT was DSpace.

Adding new features required most of the time and effort. For adding many documents fast, batch uploading was implemented. Along with automatic metadata extraction from public databases, storing new digital content got relatively easy and convenient.

Commenting and recommendation add-in was integrated into DSpace. For recommendation, two algorithms – "Social Information Filtering" and "Item-Based Top-N Algorithm" were used.

Kirjandus

[BB01] Michael J. Bass and Margret Branschofsky. DSpace at MIT: meeting the challenges. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, page 468, New York, NY, USA, 2001. ACM Press.

[BDM03] Amy Brand, Frank Daly, and Barbara Meyers. *Metadata Demystified*, 2003.

[BII] BIIT. `http://www.bioinf.ebc.ee/`.

[CRN] CRNI Handle System. `http://www.handle.net`.

[DK04] Mukund Deshpande and George Karypis. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, 22(1):143–177, 2004.

[DSD] DSpace functional documentation. `http://dspace.org/technology/system-docs/functional.html`.

[DSP] DSpace. `http://www.dspace.org`.

[Dub03] Dublin Core Metadata Initiative. *Evolving Metadata Needs for an Institutional Repository: MIT's DSpace*, 2003.

[FED] Fedora. `http://www.fedora.info`.

[GRE] A brief history of the greenstone digital library software. `http://greenstonewiki.cs.waikato.ac.nz/wiki/gsdoc/others/Greenstone_history.htm`.

[Han04] Yan Han. Digital content management: the search for a content management system. *Library Hi Tech*, 22:355–365, 12 2004.

[HK06] Hans-Werner Hilse and Jochen Kothe. Implementing persistent identifiers. Technical report, Research and Development Department of the Goettingen State and University Library, 11 2006.

[KSCS04] Anoop Kumar, Ranjani Saigal, Robert Chavez, and Nikolai Schwertner. Architecting an extensible digital repository. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 2–10, New York, NY, USA, 2004. ACM Press.

[LCB+06] Huajing Li, Isaac G. Councill, Levent Bolelli, Ding Zhou, Yang Song, Wang-Chien Lee, Anand Sivasubramaniam, and C. Lee Giles. CiteSeer: a scalable autonomous scientific digital library. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*, page 18, New York, NY, USA, 2006. ACM Press.

[LUC] Lucene. `http://lucene.apache.org`.

[MIN] Dspace'i kommenteerimise ja soovitamise funktsionaalsus. `http://dspace-dev.dsi.uminho.pt:8080/en/research_about.jsp`.

[SUB] DSpace'i seadistatav dokumentide lisamise protsess. `http://wiki.dspace.org/index.php/ConfigurableSubmissionSystem`.

[SWZ05] Gopalan Sivathanu, Charles P. Wright, and Erez Zadok. Ensuring data integrity in storage: techniques and applications. In *StorageSS '05: Proceedings of the 2005 ACM workshop on Storage security and survivability*, pages 26–36, New York, NY, USA, 2005. ACM Press.

[WBB+06] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y. Geer, Yuri Kapustin, Oleg Khovayko, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Vadim Miller, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 35:D5–D12, 12 2006.

[WBBM00] Ian H. Witten, Stefan J. Boddie, David Bainbridge, and Rodger J. McNab. Greenstone: a comprehensive open-source digital library software system. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 113–121, New York, NY, USA, 2000. ACM Press.

[Whi01] John White. ACM opens portal. *Communications of the ACM*, 44(7):14–ff, 2001.

Lisad

Lisa 1. OAI-PMH päringu XML kujul vastus

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
                             http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2007-02-25T12:33:55Z</responseDate>
  <request verb="GetRecord" identifier="oai:pubmedcentral.nih.gov:137770"
           metadataPrefix="oai_dc">
    http://www.pubmedcentral.nih.gov/oai/oai.cgi
  </request>
  <GetRecord>
    <record>
      <header>
        <identifier>oai:pubmedcentral.nih.gov:137770</identifier>
        <datestamp>2003-11-26</datestamp>
        <setSpec>pnas</setSpec>
      </header>
      <metadata>
        <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
                   xmlns:dc="http://purl.org/dc/elements/1.1/"
                   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
                                       http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
          <dc:title>
            Progesterone receptor knockout mice have an improved
            glucose homeostasis secondary to ß-cell proliferation
          </dc:title>
          <dc:creator>Picard, Frédéric</dc:creator>
          <dc:creator>Wanatabe, Mitsuhiro</dc:creator>
          <dc:creator>Schoonjans, Kristina</dc:creator>
          <dc:creator>Lydon, John</dc:creator>
          <dc:creator>O'Malley, Bert W.</dc:creator>
          <dc:creator>Auwerx, Johan</dc:creator>
          <dc:subject>Biological Sciences</dc:subject>
          <dc:description>
            Gestational diabetes coincides with elevated circulating progesterone levels.
            We show that progesterone accelerates the progression of diabetes in female
            db/db mice. In contrast, RU486, an antagonist of the progesterone receptor
            (PR), reduces blood glucose levels in both female WT and db/db mice.
            Furthermore, female, but not male, PR-/- mice had lower fasting glycemia
            than PR+/+ mice and showed higher insulin levels on glucose injection.
            Pancreatic islets from female PR-/- mice were larger and secreted more
            insulin consequent to an increase in ß-cell mass due to an increase in
            ß-cell proliferation. These findings demonstrate an important role of
            progesterone signaling in insulin release and pancreatic function and
            suggest that it affects the susceptibility to diabetes.
          </dc:description>
          <dc:publisher>National Academy of Sciences</dc:publisher>
          <dc:identifier>
```

```
    http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137770
      </dc:identifier>
          <dc:type>Text</dc:type>
          <dc:language>en</dc:language>
        </oai_dc:dc>
      </metadata>
    </record>
  </GetRecord>
</OAI-PMH>
```

Lisa 2. Pubmed veebiteenuse *efetch* poolt tagastatav XML-kujul metainfo

```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st January 2007//EN"
    "http://www.ncbi.nlm.nih.gov/entrez/query/DTD/pubmed_070101.dtd">
<PubmedArticleSet>
  <PubmedArticle>
    <MedlineCitation Owner="NLM" Status="MEDLINE">
      <PMID>1174891</PMID>
      <DateCreated>
      <Year>1975</Year>
      <Month>12</Month>
      <Day>18</Day>
      </DateCreated>
      <DateCompleted>
        <Year>1975</Year>
        <Month>12</Month>
        <Day>18</Day>
      </DateCompleted>
      <DateRevised>
        <Year>2004</Year>
        <Month>11</Month>
        <Day>17</Day>
      </DateRevised>
      <Article PubModel="Print">
        <Journal>
          <ISSN IssnType="Print">0007-1447</ISSN>
          <JournalIssue CitedMedium="Print">
            <Volume>3</Volume>
            <Issue>5986</Issue>
            <PubDate>
              <Year>1975</Year>
              <Month>Sep</Month>
              <Day>27</Day>
            </PubDate>
          </JournalIssue>
          <Title>British medical journal</Title>
        </Journal>
        <ArticleTitle>
          Letter: Distribution of oral contraceptives.
        </ArticleTitle>
        <Pagination>
          <MedlinePgn>766</MedlinePgn>
        </Pagination>
        <AuthorList CompleteYN="Y">
          <Author ValidYN="Y">
            <LastName>Chisholm</LastName>
            <ForeName>N</ForeName>
            <Initials>N</Initials>
```

```
      </Author>
    </AuthorList>
    <Language>eng</Language>
    <PublicationTypeList>
      <PublicationType>Journal Article</PublicationType>
    </PublicationTypeList>
  </Article>
  <MedlineJournalInfo>
    <Country>ENGLAND</Country>
    <MedlineTA>Br Med J</MedlineTA>
    <NlmUniqueID>0372673</NlmUniqueID>
  </MedlineJournalInfo>
  <ChemicalList>
    <Chemical>
      <RegistryNumber>0</RegistryNumber>
      <NameOfSubstance>Contraceptives, Oral</NameOfSubstance>
    </Chemical>
  </ChemicalList>
  <CitationSubset>AIM</CitationSubset>
  <CitationSubset>IM</CitationSubset>
  <MeshHeadingList>
    <MeshHeading>
      <DescriptorName MajorTopicYN="Y">
        Contraceptives, Oral
      </DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N">
        Female
      </DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N">
        Humans
      </DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="Y">
        Nurses
      </DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="Y">
        Prescriptions, Drug
      </DescriptorName>
    </MeshHeading>
  </MeshHeadingList>
</MedlineCitation>
<PubmedData>
  <History>
    <PubMedPubDate PubStatus="pubmed">
```

```
          <Year>1975</Year>
          <Month>9</Month>
          <Day>27</Day>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="medline">
          <Year>1975</Year>
          <Month>9</Month>
          <Day>27</Day>
          <Hour>0</Hour>
          <Minute>1</Minute>
        </PubMedPubDate>
      </History>
      <PublicationStatus>ppublish</PublicationStatus>
      <ArticleIdList>
        <ArticleId IdType="pubmed">1174891</ArticleId>
      </ArticleIdList>
    </PubmedData>
  </PubmedArticle>
</PubmedArticleSet>
```

Lisa 3. DSpace töövooprotsesside defineerimine

```
<?xml version="1.0"?>
<!DOCTYPE item-submission>
 <submission-map>
   <!-- By default, all collections use the "Full submission process"
        submission-process -->
   <name-map collection-handle="default">
      <submission-process id="1"/>
      <submission-process id="2"/>
   </name-map>
   <name-map collection-handle="123456789/4">
      <submission-process id="2"/>
   </name-map>
 </submission-map>
 <step-definitions>
     <step id="collection">
       <heading></heading>
       <class-name>org.dspace.submit.step.SelectCollectionStep</class-name>
       <workflow-editable>false</workflow-editable>
     </step>
 </step-definitions>
 <submission-definitions>
   <!-- This "Full submission process" process defines the DEFAULT item
       submission process -->
   <submission-process id="1" name="Full submission process">
       <!--Step 1 will be to gather initial information-->
      <step>
        <heading>jsp.submit.progressbar.describe</heading>
        <class-name>org.dspace.submit.step.InitialQuestionsStep</class-name>
        <review-jsp>/submit/review-init.jsp</review-jsp>
        <workflow-editable>true</workflow-editable>
      </step>
      <!--Step 2 will be to Describe the item.-->
      <step>
        <heading>jsp.submit.progressbar.describe</heading>
        <class-name>org.dspace.submit.step.DescribeStep</class-name>
        <review-jsp>/submit/review-metadata.jsp</review-jsp>
        <workflow-editable>true</workflow-editable>
      </step>
      <!--Step 3 will be to Upload the item-->
      <step>
        <heading>jsp.submit.progressbar.upload</heading>
        <class-name>org.dspace.submit.step.UploadStep</class-name>
        <review-jsp>/submit/review-upload.jsp</review-jsp>
        <workflow-editable>true</workflow-editable>
      </step>
      <!--Step 4 will be to Verify everything -->
      <step>
        <heading>jsp.submit.progressbar.verify</heading>
        <class-name>org.dspace.submit.step.VerifyStep</class-name>
```

```
        <workflow-editable>true</workflow-editable>
       </step>
       <!--Step 5 will be to Sign off on the License-->
       <step>
         <heading>jsp.submit.progressbar.license</heading>
         <class-name>org.dspace.submit.step.LicenseStep</class-name>
         <workflow-editable>false</workflow-editable>
      </step>
    </submission-process>
    <!-- Submission process where metadata is extracted from predefined providers -->
    <submission-process id="2" name="Automatic Metadata Extraction">
      <!-- Step 1 will be to select the database for metadata extraction and to
           insert identifier of the item in the database -->
      <step>
        <heading>jsp.submit.progressbar.describe</heading>
        <class-name>org.dspace.submit.step.SelectMetadataProviderStep</class-name>
        <review-jsp>/submit/review-extracted-metadata.jsp</review-jsp>
        <workflow-editable>true</workflow-editable>
      </step>
      <!-- Step 2 will be to Upload the item-->
      <step>
        <heading>jsp.submit.progressbar.upload</heading>
        <class-name>org.dspace.submit.step.UploadStep</class-name>
        <review-jsp>/submit/review-upload.jsp</review-jsp>
        <workflow-editable>true</workflow-editable>
      </step>
      <!--Step 3 will be to Verify everything -->
      <step>
        <heading>jsp.submit.progressbar.verify</heading>
        <class-name>org.dspace.submit.step.VerifyStep</class-name>
        <workflow-editable>true</workflow-editable>
      </step>
    </submission-process>
  </submission-definitions>
</item-submission>
```