TALLINNA ÜLIKOOL

Informaatika Instituut

Kaur Männiko

XML-ANDMEBAASI KASUTAMINE TERMINIHALDUSSÜSTEEMIS

Bakalaureusetöö

Juhendaja: Jaagup Kippar

Autor: ……..... „…“ ……. 2008

Juhendaja: ……..…….. „…“ ……. 2008

Instituudi juhataja: …….. „…“ ……. 2008

Tallinn 2008

Sisukord

Sissejuhatus

Käesolev töö tegeleb põhiliselt temaatikaga, mis kuulub terminoloogia ja infotehnoloogia valdkonna ühisossa. Töö teema on autor valinud seoses oma tööga Eesti Terminoloogia Keskuses (ETK), mis on üks Eesti Keele Instituudi (EKI) allüksusi. Infotehnoloogia (IT) annab terminoloogiatööks vajalikud vahendid info kogumise, hoidmise, töötluse ja edastamise jaoks. Töö käigus antakse ülevaade mõnedest terminoloogia valdkonna IT-vajadustest ning püütakse leida neile lahendusi.

Terminoloogiatöös ette tulevad IT-vajadused on suhteliselt spetsiifilised ja ebastandardsed. Kasutatakse palju erinevaid standardeid, formaate, andmebaase, tarkvarasüsteeme ja platvorme, mistõttu tekib sageli vajadus nende vahel andmeid teisendada, võrrelda või analüüsida. Sellise töö jaoks ei ole standardseid lahendusi, tihti tuleb improviseerida, otsustada kas ja millise töö osa jaoks programmeerida rakendus, kasutada olemasolevat tarkvara või teha hoopis mingi osa tööst käsitsi.

Üheks probleemiks ETKs on kõiki vajadusi rahuldava terminihaldussüsteemi puudumine. Töö püüab uurida, missugused on need vajadused, ja kas leidub juba olemasolevat tarkvara, mis neid vajadusi võiks rahulda, või oleks mõttekas luua midagi uut.

Lahenduste otsimise käigus keskendutakse ühe üsna uue ja kaugeltki mitte lõpuni välja arenenud tehnoloogia kasutusvõimaluste uurimisele, milleks on XML-andmebaasid. Töö annab ülevaate, mida need endast kujutavad ja miks need süsteemid võiksid olla huvipakkuvad terminihalduse kontekstis. Uuritakse XML-andmebaasisüsteemide kasutusvõimalusi ja kasutamise mõttekust terminihaldussüsteemides.

Täpsemalt on käesoleva bakalaureusetöö eesmärgid järgmised:

- anda ülevaade mõnedest hetkel olemasolevatest ja kasutatavatest terminihaldussüsteemidest, nende funktsionaalsustest ja kasutatavatest standarditest;
- anda ülevaade mõnedest XML-andmebaasisüsteemidest;
- anda ülevaade tarkvara funktsionaalsustest, mida terminoloogiatöös vajatakse ning seeläbi luua terminoloogiatöö tarkvara (terminihalduskeskkonna)

spetsifikatsioon. Ühtlasi välja pakkuda tehnoloogiaid, mida kasutada ja produkte, mida kohandada, et luua ETK jaoks vajalik uus terminihaldussüsteem ning eesti terminoloogia edendamiseks vajalik avalik terminihalduskeskkond [Sutrop07_2];

- võimalusel realiseerida prototüübid mõnedest kavandatava terminihalduskeskkonna komponentidest;

Töö koosneb neljast peatükist:

Esimeses peatükis antakse ülevaade terminoloogiatööst üldiselt, olemasolevatest terminihaldussüsteemidest ja põhiprintsiipidest terminiandmebaaside ehk terminibaaside koostamisel.

Teises peatükis tutvustatakse mõningaid XML-andmebaasisüsteeme ja uuritakse nende põhilisi omadusi ning kasutamisvõimalusi terminibaaside tarbeks.

Kolmandas peatükis antakse ülevaade kavandatavast universaalsest avatud terminihaldussüsteemist ja kirjeldatakse selle põhilisi nõudeid ja ülesehitust.

Neljandas peatükis kirjeldatakse XML-andmebaaside praktilist kasutamist loodud prototüüpide näitel.

1 Terminihaldussüsteemid

1.1 Terminihalduse põhimõtted

1.1.1 Millega tegeleb terminoloogia

Sõna terminoloogia kasutatakse peamiselt kahes tähenduses:

1) oskussõnavara;
2) oskussõnaõpetus;

Terminoloogiatöö on mõistete ja nende tähiste süstemaatilise kogumise, kirjeldamise, töötluse ja esitusega seotud töö [ISO 1087-1].

1.1.2 Terminibaas ja sõnastik

Terminibaas on terminiandmeid sisaldav andmebaas. Terminibaas on reeglina mõistepõhine. Mõiste on teadmusüksus, mille moodustab ühene tunnuste kombinatsioon. Terminisõnastik on ühe või mitme valdkonna mõistete või tähistega seotud teavet esitavate terminiartiklite kogu. Terminiartikkel on terminiandmekogu osa, mis sisaldab ühe mõistega seotud terminiandmeid. [ISO 1087-1]

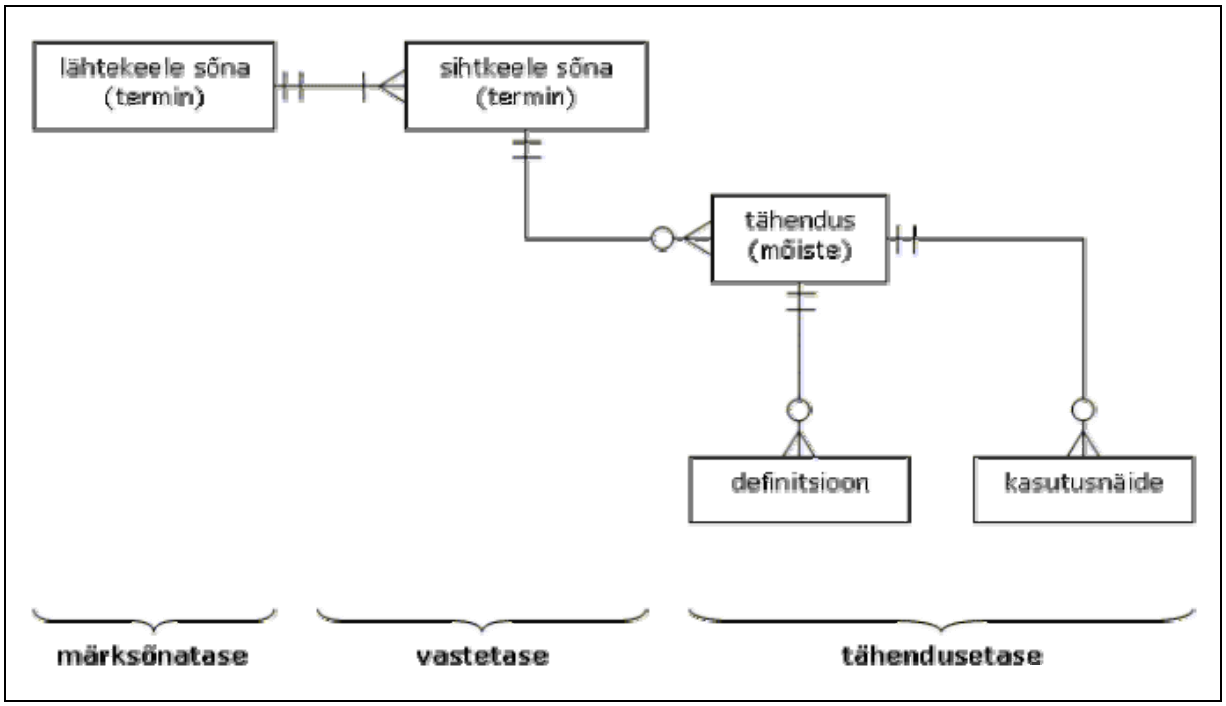
Edaspidi kasutame sõna terminikirje samas tähenduses, mis terminiartikkel.

Mõistete ja terminite vahekorda illustreerib ja aitab hästi ette kujutada VisualThesaurus.com [visthes], mis visualiseerib sõnade ja tähenduste vahelisi seoseid.

Terminibaasi võivad moodustada mõiste-, allika- ja valdkonnakirjed.

Terminibaasi koostamisaegne kuju ja avaldamisaegne kuju ei ole omavahel seotud. Kogu terminibaasis olevat infot ei pea sõnastiku lõppkasutajale esitama, sest terminibaasis võib olla sellist infot, mis on vajalik ja oluline eelkõige sõnaraamatu koostajale [Lauk05, lk 12]. Lõpptootena valmivas sõnastikus võib info paigutada kuidas tahes, näiteks tähestikjärjekorras. Sõnastikku ei jõua mitte mõistelise andmebaasi vorm, vaid sisuline kvaliteet [EOS, lk 32].

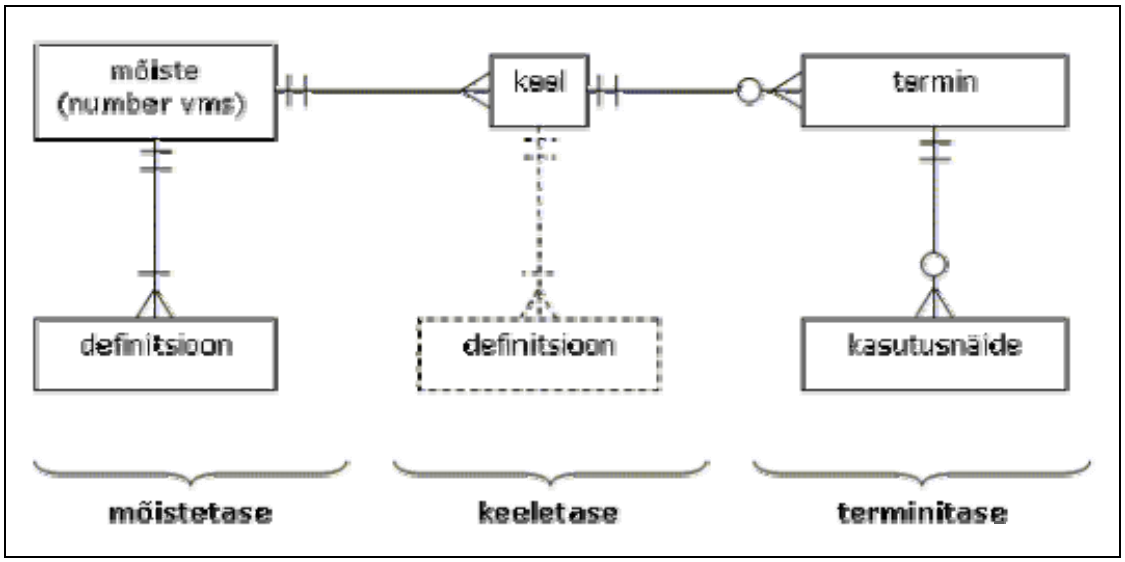
Järgnevalt on esitatud sõnapõhise andmebaasi e sõnastiku andmemudel (vt Joonis 1).



Joonis 1. Lihtsustatud sõnastiku andmemudel

Oskussõnastike kõrge kvaliteedi tagab nende loomisel eelkõige mõistepõhine (mitte sõnapõhine) lähenemine. Erinevalt sõnapõhiselt koostatud sõnastikust asuvad mõistelisena koostatava sõnastiku andmebaasis homonüümid ja polüseemid eri kirjetes, mis lubab neile täpselt viidata, kasutajale rohkem infot pakkuda ja saavutada sõnapõhisest tunduvalt täpsem vastendus. [EOS, lk 32] Just täpsusaste ongi üks olulisemaid erinevusi üldkeele ja oskuskeele vahel.

Mõistepõhisel andmemudelil (vt Joonis 2) põhinevat terminibaasi on võimalik kerge vaevaga muuta sõnastikuks ja hiljem terminibaasist saadud sõnastikku tagasi terminibaasi kujule teisendada. Vajadusel saab mõistepõhiselt koostatud terminibaasi teisendada ka teistsuguseks (nt algselt inglise-eesti suunalise eesti-inglise suunaliseks) sõnastikuks. Sõnapõhiselt koostatud sõnastikku aga samal põhimõttel teisendada ei saa. Levinumad vead, mis tekivad erialasõnastike sõnapõhisel koostamisel on nt sünonüümivastuolud, viidatavate kirjete puudumine, süsteemivabalt esitatud sugulasmõisted ja sassiläinud mõistepesad.



Joonis 2. Lihtsustatud terminibaasi andmemudel

1.1.3 Mis on terminihaldussüsteem

Terminihaldussüsteem (*terminology management system*) (edaspidi THS) on süsteem, mis võimaldab hallata terminibaase – neid luua, muuta ja esitada; terminibaasis terminiartikleid lisada, muuta, kustutada, filtreerida, teisendada; otsida termineid, termini vasteid ja muid terminiandmeid jne. On olemas iseseisvad ja tõlkemälusüsteemide koosseisu kuuluvad THS-id. Mõned THS-id kuuluvad ka mõlemasse kategooriasse korraga.

THS-e on väljatöötatud palju, nii kommertsiaalseid kui ka vabavaralisi avatud süsteeme. Üks tuntumaid kommertsiaalseid THS-e on näiteks SDL Multiterm. Terminihaldussüsteeme leidub väga erineva funktsionaalsuse ning sihtotstarbega, näiteks meditsiini valdkonna jaoks on välja töötatud mitmeid lahendusi. Mõned neist on ka avatud tarkvara, kuid oma spetsiifika tõttu pole leidnud kasutust teistes valdkondades.

1.1.4 Andmeformaadid ja standardid

Erinevad terminihaldussüsteemid kasutavad terminiandmete salvestamiseks erinevaid andmeformaate. Kommertsiaalsetel süsteemidel on tavaliselt mingi sisemine kinnine andmebaasiformaat, kuid nad võimaldavad andmete eksporti ja importi rohkematest või vähematest üldkasutatavatest formaatidest, nagu lihttekst, tabulaatoriga eraldatud tekst

(*tab delimited text*), CSV (*comma-separated values*), *MS Excel* vms. Viimased ei ole piisavalt täpsed, et esitada kogu andmebaasis olevat informatsiooni struktuurses vormis. Tootjaspetsiifiliste formaatide puuduseks on see, et neid ei toeta teised süsteemid ja vajadusel tuleb programmeerida konvertereid, mis ei ole aga tihti triviaalne ülesanne, kuna andmemudelid erinevates süsteemides võivad olla erinevad. Võib juhtuda, et tuleb leppida mõningase andmekaoga konverteerimisel.

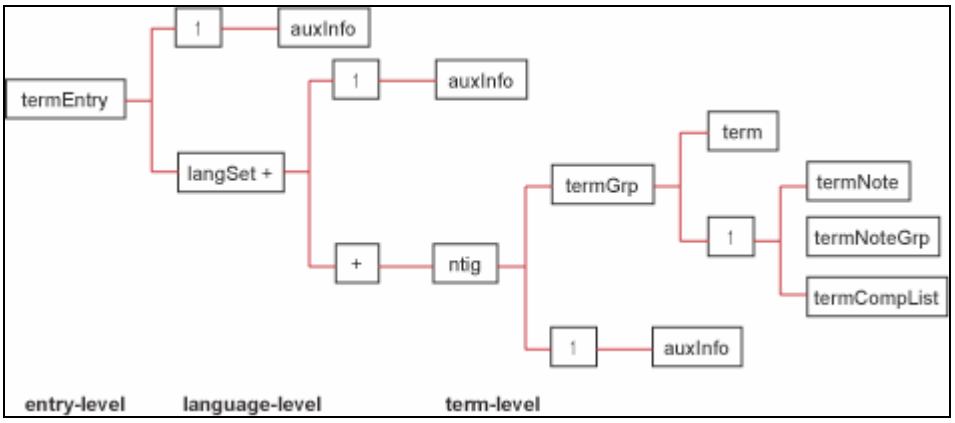
On olemas aga ka spetsiaalseid avatud formaate just terminiandmete jaoks. Uuemad neist on reeglina XML-põhised, varem on loodud ka mitte XML-põhiseid standardeid. Tootjate tugi sellistele standarditele on aga seni olnud suhteliselt tagasihoidlik [Seewald-Heeg06]. Mitmed sellised standardid jäänud laiema levikuta, kuna paljud THS-de tootjad ei ole nende standardite tuge oma programmidele lisanud. Arvata on, et see olukord peab lähemal ajal muutuma, sest vajadus terminiandmete vahetamise järele on järjest kasvav. Uute standardite väljatöötamisest ei ole õnneks loobutud ja järgnevalt antaksegi ülevaade selle ala viimastest arengutest.

Nagu p 1.1.2 kirjeldatud, on terminiartikkel hierarhilise struktuuriga ja sobib seega hästi XML-kujul esitamiseks, kuna ka XML-dokument on hierarhilise struktuuriga. XML võimaldab ilma mingite kadudeta esitada igasuguse struktuuriga andmeid ja see teeb temast pea alati parima lahenduse.

Terminiandmete vahetamiseks eri programmide vahel on välja töötatud mitmeid XML-il põhinevad standardeid, näiteks:

- XML MARTIF on mõistelise andmemudeli esitusformaat;

- TBX (*Term Base eXchange*) on suhteliselt populaarne avatud standard, mida kasutatakse näiteks ka Keeleveebis avaldatud sõnastike ettevalmistusprotsessis [Kaalep07];



Joonis 3. TBX terminikirje mudel

- Multiterm XML. SDL Multitermi poolt kasutatav mõistepõhine terminikirje formaat. Täpsemalt on see ekspordi ja impordi formaat, Multitermi sisemine formaat (nimetatakse MTF) on sama struktuuriga, kuid lühendatud märgendi nimedega. Kasutatakse muuhulgas ka ETKs. Sarnaneb struktuurilt mõneti TBX-le;
- OLIF (*Open Lexicon Interchange Format*) on avatud standard nii sõnastikulise kui ka terminiinfo vahetamiseks;
- XLT (*XML representation of Lexicons and Terminologies*) on XML-põhiste formaatide raamistik, mis põhineb TBX ja OLIF standarditel ja mis võimaldab esitada nii sõnastikke kui ka terminibaase.

1.1.5 Terminoloogiatöö Eestis

Seda, kes ja kuidas tegelevad Eestis terminoloogiatööga, on uurinud Eesti Terminoloogia Ühing (ETER). Eesti oskuskeeleseisundi uuringust selgub, et üha rohkem soovitakse koostada elektroonilisi terminibaase ja sõnastikke, eriti aladel, kus uusi termineid ja värsket infot järjest juurde tuleb [EOS, lk 51]. Uuringu käigus korraldati küsitlus, mille tulemused on esitatud ka andmebaasina ETERi kodulehel [ETER].

1.1.5.1 Eesti Terminoloogia Keskus

Eestikeelse terminoloogia arendamiseks loodi 2006 aasta alguses Eesti Terminoloogia Keskus (ETK), mis tegutseb Eesti Keele Instituudi juures.

Terminoloogiakeskus tegeleb mitmete terminoloogiaprojektidega tehes koostööd erinevate valdkondade esindajatega. Näitena võib välja tuua sõjanduse ning julgeoleku- ja kaitsepoliitika terminoloogia arendamise projekti, mille raames arendatakse terminibaasi Militerm. ETK haldab ka Eesti mahukaimat terminibaasi Esterm, millest antakse ülevaade punktis 1.1.5.2. Esterm on saadud nö päranduseks Eesti Õiguskeele Keskuselt (EÕKK), mis likvideeriti mõned aastad peale Eesti liitumist Euroopa Liiduga.

1.1.5.2 Terminibaas Esterm

Esterm on tüüpiline mõistepõhine terminibaas, kus ühes kirjes esinevad terminid on samatähenduslikud ning tähistavad ainult ühte mõistet. Seda põhimõtet on püütud järgida niivõrd, kuivõrd seda võimaldab terminoloogiatöö tõlkekeskkonnas. Paraku ei ole alati ühes kultuurikontekstis esinev reaalsus täpselt kattuv teises kultuurikeskkonnas esinevaga; mõiste võib ka sihtkeeles täiesti puududa. Vaatamata sellele tuleb lähtetekst võimalikult täpselt tõlkida ning edasi anda seal esinev terminoloogia ka sihtkeelses tekstis. Vaste valikul saab sageli määravaks kontekst, kus terminit kasutatakse. Seetõttu on iga originaaltermini juurde lisatud kontekst, kus ta konkreetsel juhul esines, ja võimalusel definitsioon originaalkeeles.

Estermis on talletatud Eesti Õiguskeele Keskuses tehtud terminoloogiatöö tulemused - üle 50 000 mõiste inglise ja eesti keeles. Kanada ekspertide abiga ISO standarditele vastavaks kujundatuna on see terminibaas leidnud tunnustust terminoloogiatöö tegijate poolt üle maailma. Päevas tehakse EÕKK kodulehel terminibaasis üle 10 000 päringu, mis näitab aktiivset kasutust kõigi vähegi tõlkimisega kokkupuutuvate inimeste poolt nii Eestis kui välismaal.

EÕKK tegeles põhiliselt Euroopa Liidu seadusandluse tõlkimisega eesti keelde ja Eesti seaduste tõlkimisega inglise keelde. Kuna tõlkimist vajavaid seadusi oli väga palju ja tõlked vajasid täpset terminoloogiat, tegeleti EÕKKs terminoloogiatööga professionaalsel tasemel ja ilmselt üldse kõige suuremas mahus Eestis. Selle tulemusel tekkis ka kõige mahukam terminibaas Eestis. Selline maht ja töökorraldus asetasid küllaltki suured nõudmised terminihaldussüsteemile. Töötati välja FoxPro ja Oracle andmebaasisüsteemil põhinev spetsiaalne terminihaldustarkvara. Arendati välja ka

koondotsinguvõimalus terminibaasist, dokumentide täistekstandmebaasist ning tõlkemäludest (tõlgitud dokumentide paralleelkorpusest). Eesti keeles on tõlkemälusid käsitlenud T. Kuub oma bakalaureusetöös [Kuub02] ja T. Toova oma proseminaritöös [Toova04].

Terminihaldussüsteem, millesse kunagi panustati üle miljoni krooni, on tänaseks iganenud. Edasiarendus jäi pooleli ja haldamine oli suhteliselt keeruline ning nõudis eriväljaõppega IT-personali pidevat hoolt. 2005 loobuti süsteemi kasutamisest ja mindi üle kommertsiaalsele terminihaldussüsteemile Trados Multiterm (nüüdse nimega SDL Multiterm), mida tutvustatakse pikemalt punktis 1.2.1.

Ühest küljest oli omatehtud süsteemist loobumine paratamatu, sest Euroopa liiduga liitumine oli ühekordne protsess ja on tänaseks seljataga. Terminoloogiatööd sellises mahus ja sellisel kujul Eestis enam ei vajata. Samas on see õpetuseks, et täna loodav tarkvara võidakse juba paari aasta pärast maha kanda, ja seda mitte ainult sellepärast, et vajadus oleks kadunud. Vajadus on jäänud, aga nõuded on muutunud.

Hetkel kasutatakse Estermi haldamiseks terminihaldussüsteemi Multiterm iX. Paraku ei vasta ka Multiterm täielikult kõikidele nõuetele, mida Esterm vajab. Näiteks on sellisteks probleemideks terminibaasi avalikustamine ja mõningad puudused funktsionaalsuses nagu piiratud otsingu- ja analüüsivõimalused.

1.1.5.3 Avalik terminihalduskeskkond

Paljudes eri valdkondades leidub inimesi, kel pole filoloogilist haridust, kuid on soov oma eriala terminoloogiat arendada ja korrastada. Paljud on nõus tegema seda ka entusiasmist, sest tunnetavad igapäevatöös korrektsete eestikeelsete terminite puudust. Selline soov on igati tervitatav ja selliseid inimesi tuleks oma püüdlustes toetada. Vastasel korral püsiks eesti keel veel mõnda aega igapäevakeelena, kuid ei oleks enam jätkusuutlik teaduskeelena.

Kvaliteetseid mõistepõhiselt koostatud oskussõnastikke ilmub vähe. Eesti oskuskeelekorralduse seisundi uuringust võib lugeda, et ei ole sugugi kindel, kas rangelt võttes võib kõiki 1996-2002 ilmunud ja uuringu aluseks olnud väidetavaid oskussõnastikke nimetada oskussõnastikeks ja neis sisalduvaid sõnu terminiteks [EOS, lk 28]. Uuringu kokkuvõttena selgub, et oskussõnastike ja terminibaaside koostamise teooria ja meetodid on praktilise töö tegijatele üsna võõrad, tihti lähtutakse muudele aladele (peamiselt üldkeele leksikograafiasse) sobivast mõtteviisist [EOS, lk 54].

Üheks lahenduseks oleks kindlasti huviliste koolitamine, millega tegeletakse näiteks ETERis. Samas ei ole sugugi lihtne leida lahendust küsimusele, kuhu asjaarmastaja võiks talletada termineid, missugust terminihaldustarkvara kasutada ja kuidas oma terminikogusid avalikustada.

Sellele probleemile püüab lahendust leida EKI, algatades projekti ühtse veebipõhise terminoloogiasõnaraamatute koostamise keskkonna loomiseks [Sutrop07_1]. Projekt on esialgu kavandatud väiksemas mahus, püüdes kohandada EKIs väljatöötatud professionaalset veebipõhist sõnastikehaldussüsteemi terminoloogiatööks.

Käesolev töö püüab laiendada eelmainitud projekti, kirjeldades nõuded XML-andmebaasil põhinevale universaalsele terminihaldussüsteemile, mille täpsem spetsifikatsioon on esitatud peatükis 0. Süsteemil on vähemalt kaks erinevat kasutajaliidest:

1) "terminoloogi liides" professionaalsele kasutajale;
2) avalik terminihalduskeskkond hobi korras terminitega tegelejale.

Üheks avaliku terminihalduskeskkonna sihtgrupiks võiksid kindlasti olla ülikoolide õppejõud ja miks mitte ka üliõpilased. Ka Eesti oskuskeelekorralduse seisundis on öeldud, et oskussõnavaraga seotud tegevus on eelkõige koondunud kõrgkoolide juurde, kus suur osa õppetöö aluseks olevatest võõrkeelsetest õppematerjalidest tuleb loenguteks eesti keeles ette valmistada. [EOS, lk 50]

Üldiselt on sõnastike loomise keskkond mõeldud kõigile, kel on huvi ja tahtmist erialasõnastikku koostada.

1.2 Olemasolevad terminihaldussüsteemid

Selles punktis on välja toodud mõningad olemasolevad terminihaldussüsteemid, mida töö autor on pidanud vajalikuks uurida. Täpsemalt, on püütud leida olemasolevaid kõige professionaalsemaid ja universaalsemaid tarkvaralahendusi. Ei saa välistada, et vaatluse alt on välja jäänud mõned üsnagi suured ja võimalusterohked lahendused, kuna paljud ettevõtted ja organisatsioonid on tunnetanud ühtse terminoloogia haldamise vajadust, siis on nad ka loonud endale sisemiseks kasutamiseks terminihalduslahendusi, mis küll enamasti lähtuvad konkreetse organisatsiooni vajadustest ja on mingil moel spetsiifilised ega sobi universaalseks kasutamiseks. Neis võib olla näiteks puudusi, mis tulenevad vähesest teoreetilisest baasist lähteülesande püstitamisel. Kui aga selliseid

süsteeme piisavalt edasi arendatakse, kiputakse välja jõudma teatud üldiste omadusteni, mis universaalsel terminibaasil kindlasti peaksid olema, nagu näiteks mõistepõhisus, universaalne ja dünaamiline kirjestruktuur, otsinguvõimalus kõikidelt väljadelt, filtrid, kasutajate ja õiguste haldamine, koostöö lahendused jne.

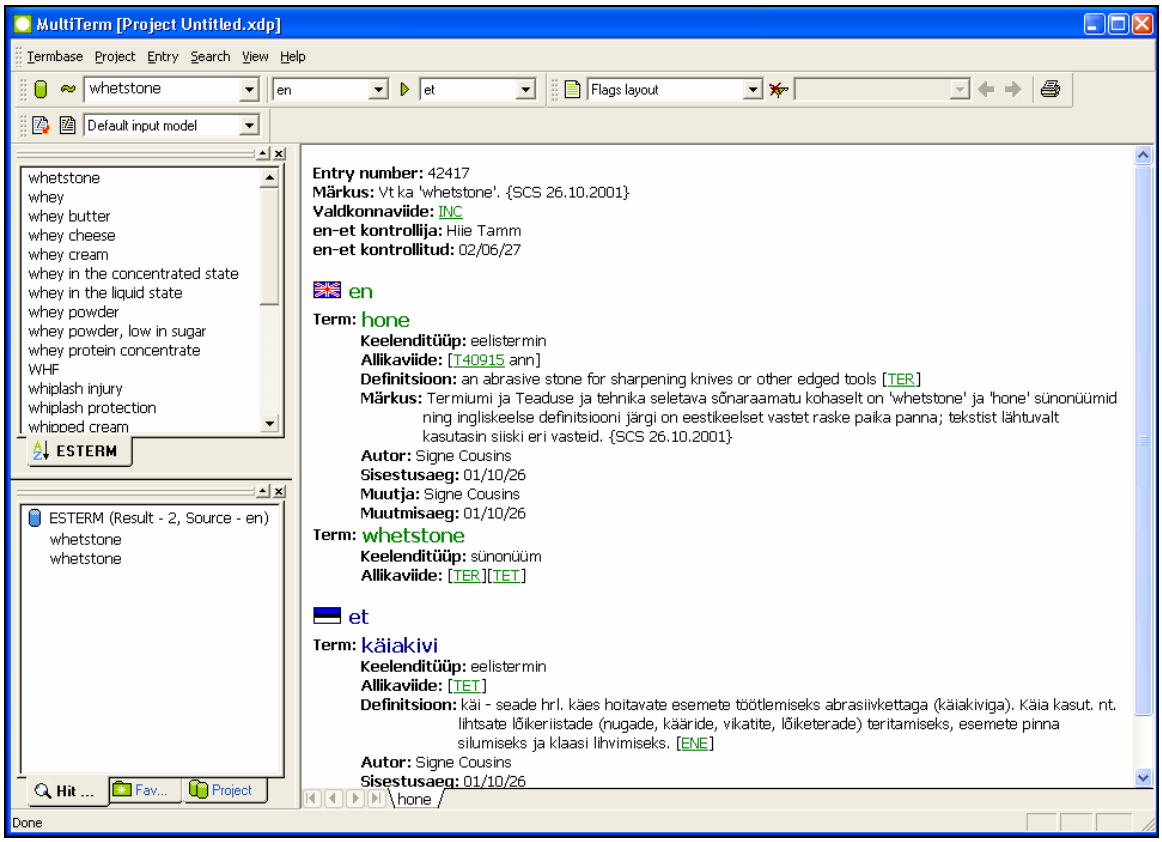
1.2.1 SDL Multiterm

SDL Multiterm on üks tuntumaid kommertsiaalseid terminoloogiahaldussüsteeme. Algselt on see välja töötatud Saksa firma Trados poolt, mis hiljem ühendati konkureeriva firmaga SDL International.

Multiterm võimaldab kasutajatel luua uusi terminibaase, lisada, muuta ja kustutada termineid, lehitseda terminibaasi, teostada otsinguid, lisada filtreid ja eksportida andmeid.

Multiterm on töölauarakendus, mida võib kasutada kui iseseisvat süsteemi, salvestades ja kasutades terminibaase lokaalselt või ka kui klient-server rakendust, mispuhul töölauarakendus on klient, mis ühendub eemal asuva serveri külge, millel jookseb Multiterm Server'i nimeline rakendus. Klient-server rakenduse puhul saab serveril asuvaid terminibaase kasutada korraga mitu kasutajat.

Multiterm on üks parimaid näiteid mõistepõhisest terminibaasist. Mõistekirjes paiknevad väljad kolmel erineval tasandil – esimesel tasandil asub mõistet puudutav info, teisel konkreetset keelt puudutav terminiinfo ja kolmandal igat konkreetset terminit puudutav info. Järgneval joonisel (vt Joonis 4) on kujutatud Multitermi versiooni „iX" kasutajaliides, parajasti avatud terminikirjega „käiakivi".

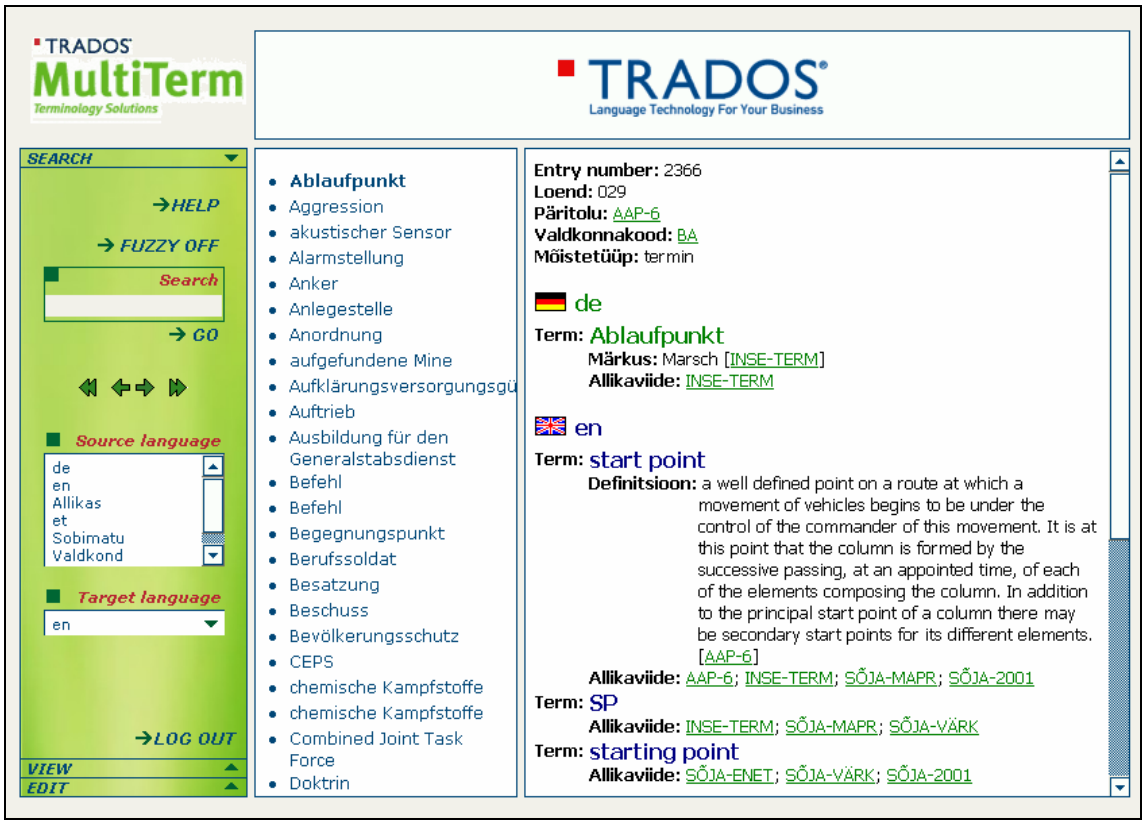


Joonis 4. Terminihaldussüsteemi Multiterm iX kasutajaliides

Samas on Multitermil ka mitmeid puudusi, olulisematest võiks esile tuua:

1) otsinguvõimaluse puudumine kõikidelt terminiartikli väljadelt;
2) viitamisvõimalus ainult terminile, mitte mõistele;
3) mittekohandatav andmete väljastus sõnastiku kujul.

Multiterm Online (vt Joonis 5) on Multitermi veebipõhine liides, mida müüakse koos Multiterm Serveriga.



Joonis 5. Multiterm Online veebiliides

Multiterm Online veebiliides ei toeta paraku muid brausereid kui Microsoft Internet Explorer, mis teeb ta kasutuskõlbmatuks suure kasutajaskonnaga terminibaaside avalikustamisel, nagu seda on näiteks Esterm. Selle tõttu on ETKs käesoleva töö autori poolt loodud nn kohandatud veebiliides (vt Joonis 6). Kohandatud veebiliides on realiseeritud kasutades Multiterm Serveri API-t (*Application Programming Interface*). Lähemalt on seda kirjeldatud töö praktilises osas p **Tõrge! Ei leia viiteallikat.**.

Joonis 6. Estermi avalik veebiliides

1.2.2 i-Term

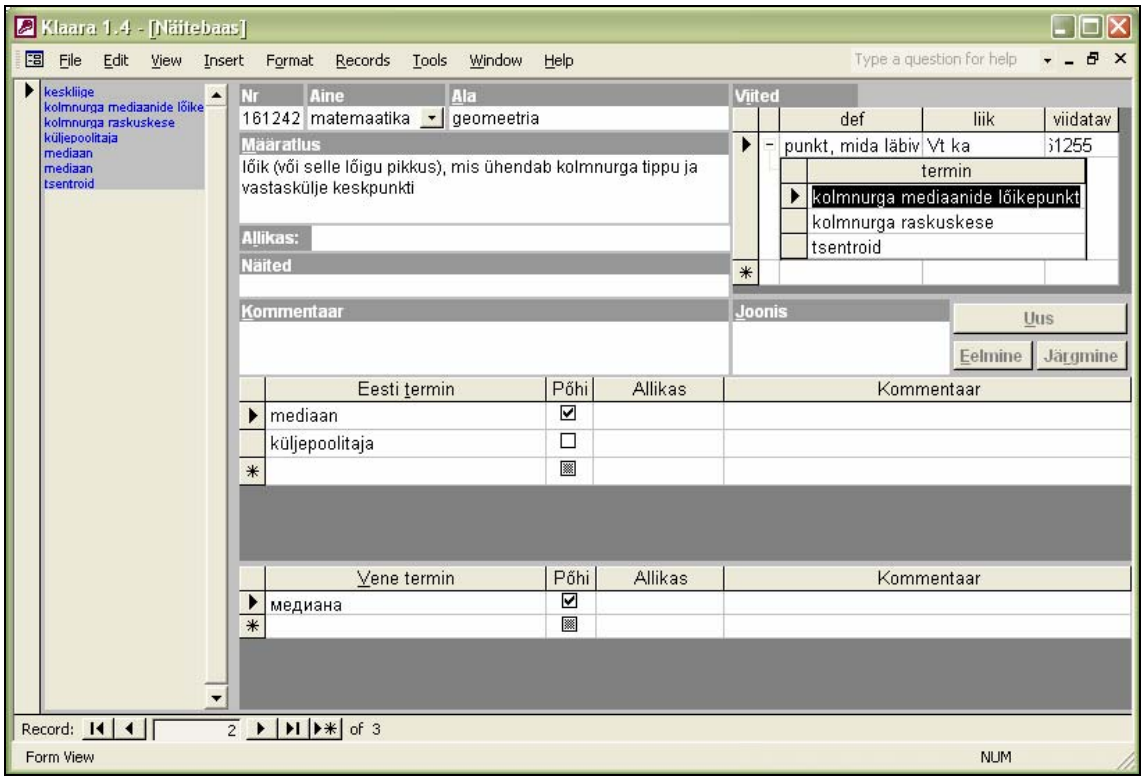
i-Term on loodud Taani riikliku terminoloogia keskuse DANTERMcentre poolt ja on tasuline veebipõhine (vt Joonis 7) terminihaldussüsteem [i-Term].

Joonis 7. i-Term'i veebiliides

Lisaks hinnale on puuduseks ka piiratud kasutajate arv. Antud süsteemiga ei olnud võimalik lähemalt tutvuda, kuna palvele saada demo-kasutaja õigused, ei vastatud.

1.2.3 Klaara

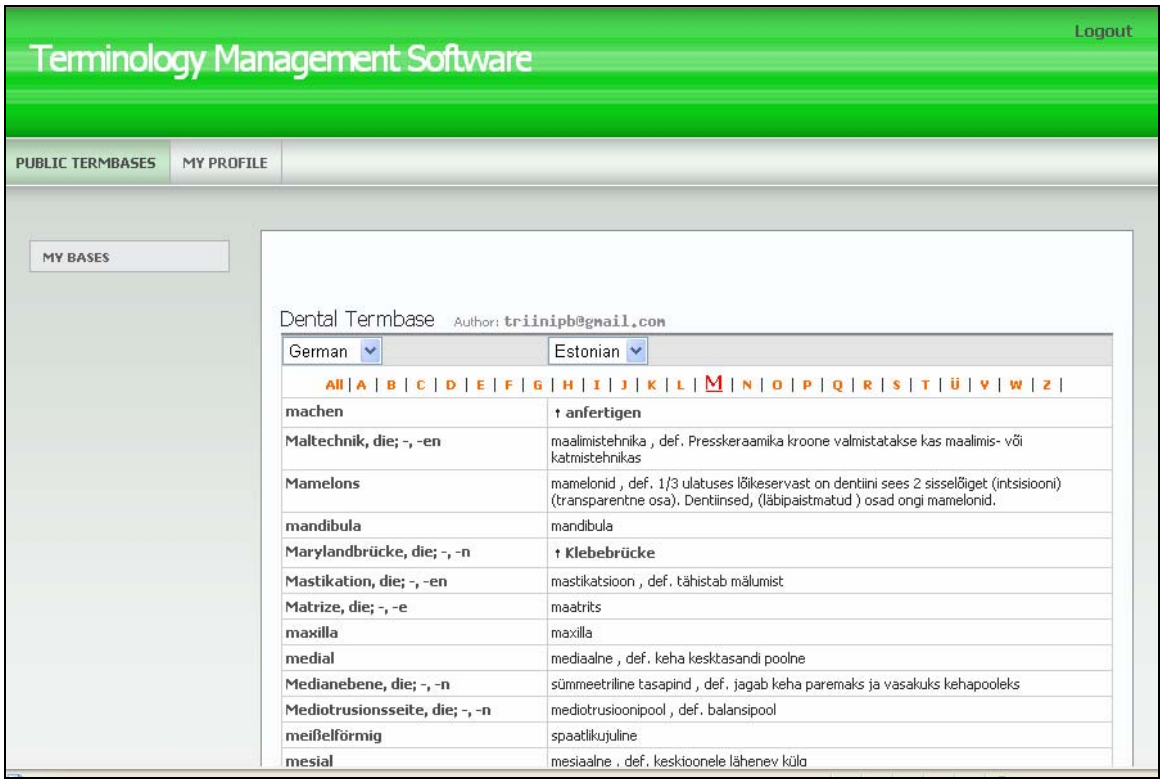
Microsoft Accessi baasil loodud tasuta andmebaasirakendus (vt Joonis 8). Klaara on mõistepõhine, realiseeritud on sõnapõhine väljund. Tegemist on töölauarakendusega, millel puudub serveri tugi ja seega ka võimalus mitme kasutaja samaaegseks tööks ühe terminibaasiga.



Joonis 8. Terminiandmebaas Klaara

1.2.4 Termbases.eu

Termbases.eu on Werkdata OÜ poolt loodud lihtne tasuta veebipõhine terminihaldusteenus [Termbases].



Joonis 9. Termbases.eu

Puudub otsinguvõimalus otsisõna sisestamise teel, võimalik on ainult sirvida termineid tähestiku järjestuses algustähe järgi.

1.2.5 Leksikograafia tööriistad

Otsides lahendusi terminiandmebaaside jaoks, tasub heita pilk ka leksikograafia valdkonnas toimuvale. Ka leksikograafid kasutavad XML-formaate täpselt samadel põhjustel, kui terminograafid. XML on universaalne ja paindlik formaat. Üldjuhul ei saa terminibaaside haldamiseks kasutada sõnastike loomise tööriistu eelkõige just neis kasutatava sõnapõhise andmemudeli tõttu. Kui aga andmed salvestatakse XML-kujul, on nii elektroonilise sõnastiku kui ka terminibaasi ühiseks jooneks see, et nad kujutavad endast hulka XML-kirjeid. Kirjed on küll erineva struktuuriga, kuid meetodid, kuidas neid salvestatakse, neist andmeid otsitakse ning töödeldakse, on universaalsed.

Sõnastike loomiseks on olemas mitmeid tarkvaralahendusi. Järgnevalt on kirjeldatud rakendust, mis on loodud Eesti Keele Instituudis sõnastike haldamiseks.

1.2.5.1 EELex

EELexi näol on tegemist AJAX (*Asynchronous JavaScript and XML*) kasutajaliidesega XML-põhise klient-server rakendusega [Loopmann07].

Senine kasutus on piirdunud sõnapõhiste andmebaasidega, kuid teoreetiliselt peaks rakendus sobima ka mõistepõhiste XML-andmebaaside jaoks. Mõistepõhise andmebaasi kasutamiseks tuleb defineerida vastav XML-skeem.

Serveri poolel on tegemist Apache veebiserveris käivitatavate Perl'i skriptidega, mis kasutavad andmete salvestamiseks harilikku XML-tekstifaili. Tekstifailist andmete otsimine ei ole eriti efektiivne lahendus ja nii on andmebaasi suurus piiratud tänapäevaste arvutite puhul umbes 20 MB XML-kujul andmetega.

Põhimõtteliselt toimub andmeotsing nii, et XPath kujul päring saadetakse klientprogrammist serveri Perl'i CGI moodulile, mis, kasutades teeki Libxml, laeb mällu kogu XML-dokumendi, millest tehakse otsing. Ja nii iga päringu puhul. Tegemist ei ole just eriti efektiivse serveriressursside kasutamisega. Võib küll hankida võimsama riistvara, kuid skaleeruvus jääb tõsiseks probleemiks juhul, kui sõnastike maht peaks kasvama või tekkima rohkem kasutajaid, mis tähendab rohkem samaaegseid päringuid.

EELexi puhul tasuks kaaluda serveri rakenduse ümberkohandamist XML-andmebaasi kasutamiseks.

Lisaks on EELexi puuduseks veel sõltuvus Microsofti platvormist, kuna kasutab MSXML parserit ning VBScripti, mis ei võimalda tema kasutajaliidest kasutada näiteks Linux keskkonnas.

2 XML-andmebaasisüsteemid

XML on universaalne hierarhiline andmeformaat, milles on võimalik esitada peaaegu igasuguseid andmeid. XML-formaat on ühelt poolt struktuurne kuid samas paindlik, võimaldades esitada nii range struktuuriga kirjeid kui ka vaba struktuuriga dokumente.

Viimastel aastatel on arenema hakanud uus valdkond – XML-andmebaasisüsteemid. XML-andmebaas on andmebaas, kus hoitakse XML-dokumente.

2.1 Liigitus

Mõned allikad eristavad peamiselt kahte liiki XML-andmebaase. Need on „XML-võimelised“ ja „päris XML-andmebaasid“ [e-teatmik].

Täpsemalt võib XML-andmebaasid jagada nelja erinevasse kategooriasse:

- Algupärased (*native*) XML-andmebaasisüsteemid.
- XML-laiendused olemasolevatele relatsioonilistele andmebaasidele.
- Hübriidandmebaasid, mis sisaldavad relatsiooniliste andmebaaside omaduste kõrval ka algupärase XML-andmebaasi omadusi. Tegemist on enamasti tuntud relatsiooniliste andmebaasitoodete edasiarendustega.
- XML-relatsioonilised teisendajad (vahevara).

T. Kaeeli on oma bakalaureusetöös [Kaeeli04] uurinud erinevaid XML-andmebaasisüsteeme, nende liike ja omadusi.

2.2 Päringukeeled ja standardid

XML-andmebaasisüsteemid kasutavad andmetele juurdepääsuks ja andmete töötlemiseks mitmesuguseid päringukeeli ja standardeid.

2.2.1 XPath

XPath on keel XML-dokumentidest alamhulkade leidmiseks. Hetkel on olemas versioonid 1.0 ja 2.0.

2.2.2 XQuery

XQuery on W3C standard ja väga paindlik ning suurte kasutusvõimalustega päringukeel. Ta võimaldab teisendada XML-dokumente, muuta elementide järjekorda, konstrueerida uusi elemente ning kombineerida omavahel andmeid mitmetest dokumentidest. XQuery sisaldab endas XPath'i (versioon 2.0), iga XPath avaldis on korrektne ka XQuery avaldisena. Lisaks XPath'ile toetab XQuery FLWOR avaldisi, mis meenutab mõneti SQL keelt.

2.2.3 XUpdate

XUpdate ehk *XML Update Language* on päringukeel XML-andmete muutmiseks. See ei ole saanud küll W3C standardiks, kuid seda toetavad mitmed süsteemid ja tööriistad.

2.2.4 XML:DB

XML:DB on tootjast sõltumatu API XML-andmebaasidele. XML:DB'd toetavad juba mitmed andmebaasid, näiteks eXist, Xindice, Sedna.

2.3 Olemasolevad produktid

XML-andmebaasisüsteeme on loodud juba üpris palju ja erinevaid. Mõnede arendamine on ka lõpetatud. Olemasolevate XML-andmebaasisüsteemide kohta leiab ülevaatlikku infot näiteks saidilt rpbourret.com [rpbourret].

Leidub avatud ja kommertsiaalseid süsteeme. Avatute puuduseks kipub olema vähene dokumentatsioon ja tugi. Kommertsiaalsete kohta leidub dokumentatsiooni rohkem.

2.4 Vajadused ja esitatavad kriteeriumid

Käesolevas punktis uuritakse mõningaid XML-andmebaase, mida saaks kasutada terminihaldussüsteemis. Kuna erinevaid tooteid on väga palju, on uurimise alt välja jäetud kõigepealt sellised süsteemid, mis mingil põhjusel kindlasti ei sobi. Kriteeriumid, millest valikul lähtutakse, on järgmised:

1) Andmemaht. Enamusel XML-andmebaasisüsteemidel on piir, kui palju nad suudavad XML-andmeid salvestada. Piiratud võib olla nii ühe dokumendi suurus, dokumentide arv, mis on salvestatud ühte kogumisse (*collection*) või ka kogumi andmemaht.

2) Jõudlus. Eriti päringu kiirus – kui kiiresti on võimalik saada andmebaasist kätte küsitud andmed. Vastuse kiirus sõltub kindlasti päringu keerukusest, vajalike indeksite olemasolust ja andmete mahust.

3) Funktsionaalsus. Vajalik on XQuery tugi, täistekstotsingu võimalused, XUpdate või analoogne andmete muutmise võimalus. Andmeid võiks olla võimalik salvestada nii XML-skeemi järgi kui ka ilma eelnevalt dokumendistruktuuri määratlemata.

4) Liidesed enamlevinud programmeerimiskeeltele ning võimalus kasutada XML-andmebaasisüsteemi serveril.

5) Hind.

Eelistatud on algupäraseid või hübriid XML-andmebaasisüsteeme, sest neil on tõenäoliselt rohkem vajalikke omadusi (näiteks XQuery tugi). Samuti peaks nende haldamine olema lihtsam.

2.5 Huvitavamate andmebaaside tutvustus ja katsetamine

Selles punktis tutvustatakse lähemalt huvitavamaid XML-andmebaase, mis esmapilgul võiksid olla potentsiaalsed kandidaadid terminihaldussüsteemis kasutamiseks vastavalt p 2.4 toodud kriteeriumitele.

2.5.1 Berkeley DB XML

Algselt firma Sleepycat poolt välja arendatud ja hiljem Oracle poolt üle võetud avatud lähtekoodiga andmebaas, millest on olemas paralleelselt relatsiooniline ja XML variant.

Tuleb toime suurte andmemahtudega, päringukiirus suhteliselt hea, aga mitte piisav. On ettenähtud kasutamiseks manustatud (*embedded*) vormis ehk programmiteegina, mitte eraldi protsessina või klient-server rakendusena.

Tähelepanuväärne on, et seda XML-andmebaasisüsteemi kasutatakse näiteks Baskimaa ülikoolis väljaarendatud sõnastikehaldussüsteemis. [Alegria]

2.5.2 eXist

eXist on avatud lähtekoodiga ja GNU LGPL litsentsiga Java-põhine algupärane XML-andmebaasisüsteem. Huvitav ja väga atraktiivne andmebaasilahendus koos sisseehitatud

veebiserveriga. Stiilipuhas näide XML-tehnoloogia rakendamisvõimalustest, kuna toetab pea kõiki sellekohaseid standardeid ja ka veebipõhise kasutajaliidese ehitamiseks ei ole tarvis kasutada muid tehnoloogiaid peale XQuery ja soovi korral XSLT. Toetab ka täistekstotsingu võimalusi XQuery laiendusena.

Kasutab mudelipõhist andmete säilitamist ja indekseid. Detailsemalt on seda eesti keeles kirjeldanud T. Kaeeli. [Kaeeli04]

Installeerimine on lihtne ja paigaldamiseks on mitmeid võimalusi, näiteks võib süsteem töötada kas iseseisvalt serverirakendusena või servleti konteineris.

Estermi suurust dokumenti ei tervelt ega kirjete kaupa ühes kogumis salvestada ei suuda. Väiksema dokumendi puhul (40 MB) on ka lihtsamad päringud mitme sekundi pikkused, mis on kindlasti kavandatava terminihaldussüsteemi jaoks liiga palju.

Ilmselt sobiks terminihaldusrakenduste loomiseks ainult tingimusel, et andmete maht ei ole suur. Päringud üle suurema andmehulga on liiga aeglased isegi indeksite kasutamise korral.

2.5.3 Sedna

Sedna on avatud lähtekoodiga ja tasuta algupärane XML-andmebaasisüsteem. Mitmetes võrdlustes on seda esile toodud just jõudluse poolest [Cuong]. Täistekstotsingu võimalus on saavutatav aga ainult teise firma toote abil, mis on tasuline ja mitte väga odav. Ka demoversiooni ei olnud võimalik testida, kuna seda lihtsalt ei saadetud.

Sedna näitas testitud vabavaralistest toodetest kõige paremat jõudlust. Lisaks võimaldab ta salvestada 300 MB dokumendi ka ühes tükis ning selle osi muuta XUpdate päringutega.

2.5.4 IBM DB2

Hübriidandmebaasisüsteemidest sai valitud IBM’i DB2. Töö kirjutamise hetkel ilmunud versioon 9.5 omab üpris head vahendite hulka XML-ga ümberkäimiseks. Andmebaasist pakutakse vabavaralist versiooni, millel on hea funktsionaalsus võrreldes konkureerivate kommertsiaalsete andmebaasisüsteemide analoogsete vabavaraliste versioonidega. Lisatav täistekstotsingu komponent on samuti tasuta.

DB2 ei loo XML-i salvestamiseks eraldi kogumeid vaid võimaldab salvestada XML-i lihtsalt tabeli lahtritesse, selleks on spetsiaalne andmetüüp „XML“. „XML“-tüüpi andmetulp kujutab endast aga analoogi algupäraste XML-andmebaaside kogumikule (*collection*), üle selle saab teha näteks XQuery päringuid. Terminikirje jaoks sobib XML-kirje hoidmine tabeli lahtris küll. Hea omadus on ka see, et sama andmebaas võimaldab salvestada ka relatsioonilisi andmeid traditsiooniliselt tabelite kujul. Kui ehitada suuremat terminihaldussüsteemi, siis on seal piisavalt relatsioonilise iseloomuga andmeid, mida ei ole mõtet XML-kujul hoida. Üks võimalus on kasutada paralleelselt XML ja relatsioonilist andmebaasi, kuid see kasvatab kohe süsteemi keerukuse taset, sest tarvis on tagada andmete kooskõlalisus eri andmebaasisüsteemide vahel.

Testiti andmebaasi vabavaralist versiooni DB2 Express-C 9.5 koos täistekstotsingu laiendusega NetSearch Extender.

Andmebaas ei toeta XUpdate päringuid, kuid analoogsed vahendid on päringukeeles siiski olemas.

2.6 XML-andmebaaside omaduste võrdlus

Järgnevalt (vt Tabel 1) on koondtabelina esitatud mõnede XML-andmebaaside olulisemad omadused.

	Avatud lähtekoodiga	Litsents	XQuery	XUpdate	XML:DB liides	Täistekst-otsingu võimalus
eXist	Jah	GNU LGPL	Jah	Jah	Jah	Jah
Berkeley DB XML	Jah	Tasuta	Jah	Jah	Ei	Ei
DB2	Ei	Jah	Jah	Ei	Ei	Jah
Sedna	Jah	Apache License 2.0	Jah	Jah	Jah	Tasuline
X-Hive/DB	Ei	tasuline	Jah	Jah	Jah	väline
Xindice	Jah	Jah	Ei	Jah	Jah	Jah

Tabel 1. XML-andmebaaside omadused

3 Nõuded terminihaldussüsteemile

Selles peatükis uuritakse, missugused oleksid ideaalsele terminihaldussüsteemile esitatavad nõuded, seda eelkõige ETK vajaduste kontekstis. Selleks on käesoleva töö autor kogunud infot ETK terminoloogidelt [ETK06_1], [ETK06_2], ETK avalike terminibaaside kasutajate tagasisidest, erialakirjandusest ja ka terminoloogiateemalistelt üritustelt.

3.1 Vajadused ja lähtealused

Pea kahe aasta jooksul, mil ETK on tegutsenud, on tehtud terminihaldusega seotud koostööprojekte eri valdkondade esindajatega. Projektide raames on esile kerkinud mitmeid vajadusi ja soove, mida praegune tehniline baas ei võimalda hästi või üldse mitte täita. Selliste vajaduste põhjal on tekkinud idee luua ühtne ja avatud terminihaldussüsteem koos selles sisalduva avaliku terminihalduskeskkonnaga.

3.2 Süsteemi funktsionaalsus

Süsteem hõlmaks endas veebikeskkondi, veebiteenuseid ja ühendusmooduleid (adaptereid) andmevahetuseks teiste süsteemidega. Süsteemi olemust võiks väljendada sõnadega "terminiandmebaaside X-tee".

Süsteemi peamine ja kasutajale otseselt nähtav osa oleks avalik internetipõhine terminihalduskeskkond. Selle eesmärk on võimaldada ilma oluliste lisakulutusteta ja pika ettevalmistusperioodita luua oskuskeelesõnastikke isikute või huvigruppide poolt, kes sellest vähegi huvitatud on. Süsteem võimaldaks optimaalsemalt organiseerida eesti terminoloogiaarendust vähese inimressursi ja väikese keelekasutajaskonna tingimustes. Lisaks aitab süsteem siduda paremini olemasolevaid terminoloogiahalduse tehnilisi ressursse.

Süsteem peab evima semantilist koosvõimet (*semantic interoperability*) teiste süsteemidega. Selleks saab süsteem pakkuda erinevaid veebiteenuseid (XML-põhiseid andmeteenuseid), ühendusstandardeid ja protokolle (nagu TBX, OLIF, MultiTerm XML jms) enda liidestamiseks teiste süsteemidega. Näiteks on võimalik andmeid vahetada SDL Multiterm Serveriga ning kasutada süsteemi Trados Workbench tõlkemälusüsteemi terminituvastusfunktsiooniga. Trados Workbench on üks enim

kasutatav tõlkemälu programm, mis võimaldab tõlkimise käigus teha päringuid ka terminibaasist.

Väljavõte riiklikust programmist „Eesti keele keeletehnoloogiline tugi (2006-2010)“: „Keeleressursid luuakse nii, et oleks tagatud nende omavaheline semantiline koostoime ning koostoime kõigi teiste infosüsteemidega. Suutlikkus teiste süsteemidega regulaarselt andmeid vahetada on muutumas süsteemide üheks peamiseks oskuseks/komponendiks üldse, ning semantiline koosvõime on selle oskuse põhituum. Semantiline koostoime ei ole absoluutne, sest ei ole võimalik kasutada ainult ühte andmebaaside struktuuri ega universaalset keelt (nt XML). Seepärast pööratakse suurt tähelepanu kasutajaliideste loomisele, mis lubab välja arendada semantiliselt koostoimivad võrgustikud ning luuakse vastavad standardid.“ [HTM05]

Iga era- või juriidiline isik saab soovi korral registreerida ennast süsteemi kasutajaks ning luua oma terminikogu. Keskkond võimaldab erialaeksperdil filoloogilist haridust omamata alustada kiirelt terminikogu koostamist ning pakub koostajale igal sammul juhiseid ja abimaterjale, aidates nii tagada talletatava terminiinfo kõrget kvaliteeti.

Loodav tarkvara võimaldab internetikeskkonnas luua 1..n keelseid terminibaase, väga lihtsa kuni väga keerulise struktuuriga. Terminibaasi struktuuri on võimalik lihtsalt muuta, ka siis, kui see on juba täidetud andmetega (nt lisada välju). Näiteks alustatakse tabeli kujul tõlkevastetega, millest areneb hiljem välja mõistepõhine, süstemaatiline ja klassifikaatori(te)ga märgendatud terminibaas. Lisades terminiartiklitele erinevaid semantilisi seoseid, on võimalik terminibaasist kujundada ka mitmekeelne tesaurus. Siiski järgivad kõik erineva struktuuriga terminibaasid teatud üldist mudelit, mis tuleneb terminibaaside koostamise põhimõtetest.

Kasutada võib mitmeid valdkonna klassifikaatoreid, näiteks Lenoch, Eurovoc, ISCED vms ja luua uusi. Klassifikaatorite vahel saab luua seoseid ning selle abil omakorda siduda eri klassifikaatoreid kasutavad terminibaasid omavahel valdkonnapõhiselt, või lähedaste valdkondade põhjal.

Valdkondade liigitusega saab siduda süsteemi kasutajad, kes on ennast registreerimisel märkinud mingi valdkonna eksperdiks või "huviliseks". Samuti on süsteemi võimalik siduda näiteks ETERi oskuskeelekorralduse andmebaasiga [ETER07], mis koondab erialaspetsialiste, nende poolt kirja pandud terminiallikaid ja muud infot. Lisaks saab luua ja valdkondadega siduda ka foorumeid.

Seega seob süsteem terminid, terminiallikad, valdkonnad, foorumid ja erialaeksperdid nii, et vajadusel on võimalik küsimuse all oleva termini juurest jõuda otse vastava eriala või sellele lähedase eriala foorumite, terminiallikate või ekspertideni (kes vabatahtlikult võivad olla nõus terminipäringutele vastama).

Kõik kirjeldatud tegevused keskkonnas on reguleeritud kasutajaõigustega. Samuti ei ole kogu informatsioon kõikidele kasutajatele nähtav. Igal keskkonna objektil (sõnastik, klassifikaator, foorum) on oma haldaja/looja, kes annab vastavale objektile juurdepääsuõigused. Näiteks erialaekspertide ja teiste kasutajate kohta saab näha ainult nii palju infot, kui nad ise on nõus avalikustama.

Erialaspetsialistidel on keskkonnas võimalik koonduda huvigruppidesse, luua foorumeid (vajadusel meililiste), kus arutleda terminite üle või osaleda virtuaalses terminoloogiakomisjoni töös, st vaadata üle ja kommenteerida väljapakutud terminiloendeid, vajadusel hääletada väljapakutud terminite poolt jne. Sotsiaalse tarkvara funktsioonid ja kommunikatsiooni võimalused peaksid olema terminihaldussüsteemi loomulik osa.

Keskkonnas saab ka luua teadmusbaase wiki kujul, mille artiklitega saab siduda terminikirjeid ja muid objekte. Põhimõtteliselt erineb wiki artikkel terminiartiklist ainult oma vabama struktuuri poolest.

Keskkonnas loodud erinevaid terminikogusid on võimalik koondada suurematesse koondterminibaasidesse ja ühisotsingusse, samas on võimalik iga terminikoguga eraldi tööd jätkata, nii et muudatused kajastuvad ka koondterminibaasis, juhul kui need on avalikustatud. Või vastupidi, soovi korral on võimalik originaalsõnastikust teha ka "virtuaalne koopia", milles tehtavad muudatused ei kajastu suuremates terminikogudes vaid on nähtavad ainult ühele kasutajale või kasutajate grupile. Vastavalt virtuaalse koopia looja soovile võivad originaalsõnastikus tehtavad muudatused kajastuda virtuaalses koopias või seal mitte kajastuda. Kui selliste paralleelversioonide puhul tekivad konfliktid (ühte ja sama andmeelementi muudetakse mõlemas versioonis), võimaldab süsteem selliseid konfliktseid kirjeid leida ja ühendada.

Keskkond võimaldab näidata terminiandmeid erinevates vaadetes nagu artiklivaade, tabelivaated, XML, trükiversiooni küljendusvaade jm, muuhulgas näiteks ka graafiliselt kujutatud mõistesüsteemid, defineerida ning kasutada erinevaid filtreid ja grupeerimisi.

Terminikogud, klassifikaatorid jt andmeobjektid võimaldavad versioonimist – kõik terminikirje ehk andmeobjekti muudatused säilivad andmebaasis ja vajadusel on võimalik muudatusi võrrelda ning tagasi võtta (nagu ka näiteks wiki artiklite puhul).

Kõiki funktsionaalsusi ei arendata välja kohe, esmalt luuakse olulisemad moodulid. Võimalusel püütakse kasutada juba loodud avatud tarkvarakomponente (foorumid, wikid). Tegemist ei oleks ühe monoliitse ja gigantse suletud tarkvaraga vaid pigem komponentide kooslusega, kus iga komponent täidab oma kindlaksmääratud rolli ja on eraldi arendatav, võimalusel iseseisvalt rakendatav ning võimalikult sõltumatu teistest.

Tarkvara litsentsi valik peaks võimaldama seda kasutada kõigil soovijatel ning soodustama arendajate kaasamist väljapoolt.

3.3 Tarkvara arhitektuur

Terminihaldussüsteemi tarkvara võib jagada andmebaasi kihiks, vahevara kihiks, teenuste kihiks ning kasutajaliideste kihiks. Kuna nii andmebaasi kiht (milleks on plaanis kasutada XML-andmebaasi) kui ka kasutajaliidese kiht (milleks on plaanis kasutada AJAX tehnoloogiat) on hetkel üsna kiirelt arenevas etapis, on väga oluline, et tarkvara arhitektuur oleks võimalikult nõrgalt seostatud (*loosely coupled*). Kaaluda tasub ka teenuspõhise arhitektuuri (*service oriented architecture*) põhimõtteid.

3.3.1 Andmebaasi kiht

Andmete salvestamiseks on plaanis kasutada XML-andmebaasi. Kuna valdkond on alles väljakujunemisjärgus ja areneb väga kiiresti, siis ei tohiks süsteemi üles ehitada ühel konkreetsel andmebaasisüsteemil põhinevaks. Kui turule tuleb mõni uus ja parem andmebaasisüsteem, peab saama sellele üle minna, ilma et süsteemi teisi osi tuleks muuta. See tähendab, et tuleb lähtuda üldistest standarditest nagu XQuery, XUpdate, XSLT, XML:DB jne. Võib arvata, et lähemas tulevikus arendab iga endast lugupidav andmebaasitootja välja toe nendele standarditele.

3.3.2 Vahevara

Vahevara sisaldab äriloogikat ja seob allteenused lõpprakendustega – samas muudab nad omavahel sõltumatuks.

3.3.2.1 Konveiertöötluse raamistik

Kuna termihalduses tekib ikka ja jälle vajadus terminiandmeid teisendada erinevate andmeformaatide vahel, filtreerida või muuta andmete struktuuri ja tihti on võimalik eristada nendes teisendustes sarnase funktsionaalsusega osi, mis enamasti asuvad ahelana järjestikku, siis oleks hea, kui neid osi saaks võimalusel korduvalt kasutada ja juba olemasolevatest osadest lihtsalt uusi teisenduste ahelaid kombineerida. Nimetame neid osi komponentideks. Sellised komponendid peaksid olema oma funktsionaalsuselt võimalikult atomaarsed, et neid saaks kasutada võimalikult paljudel juhtudel. Samuti peaksid nad asuma ühtses raamsüsteemis, mis võimaldaks kerge vaevaga moodustada neist andmetöötlusahelaid. Iga komponent selles ahelas võiks olla omaette lõim (*thread*), mis tähendaks, et teisenduste ahelast koosnev protseduur suudaks paremini ära kasutada tänapäeval juba üpris tavaliseks saanud mitmetuumalisi protsessoreid või multiprotsessorsüsteeme. Sellise ahela erinevad komponendid võivad töötada isegi erinevates füüsilistes masinates. Konveiertöötluse raamistik peaks olema ka platvormist sõltumatu.

Võimaldab vähese vaevaga kombineerida vajalikke teisendusprotseduure, mida terminibaasidega opereerimisel võib vaja minna. Näiteks ühekordsel või regulaarsel andmete teisendamisel või nõudmisel (reaalajas) päringute töötlusel. Peaks võimaldama kasutada erinevaid tehnoloogiaid: XSLT, XQuery, Java teegid jne.

XSLT kui funktsionaalne keel ei sobi alati XML-teisenduste kirjeldamiseks kõige paremini. Mõnikord on protseduurses programmeerimiskeeles vajaliku teisenduse kirjeldamine lihtsam.

Suuremate XML-failide töötlemine võib osutuda probleemseks, kuna konkreetne teisendaja võib vajada tööks mälu hulka, mis on võrdeline XML-dokumendi suurusega – st teisendaja ei oska töödelda dokumenti jadamisi, kuigi dokument koosneb kirjetest, mis ei ole töötlusprotseduuris omavahelises sõltuvuses. Selleks võimaldab raamistik dokumendi lugemisel jagada see kirjete kaupa XML-fragmentideks, mida on seejärel võimalik suunata voogudesse, mis ühendavad erinevaid teisendajaid. Voogude kasutamine võimaldab teisendajaid käitada erinevates lõimedes, kusjuures iga lõim võib joosta erineval füüsilisel protsessoril ehk protsessori tuumal või isegi erineval füüsilisel arvutil. Selline lahendus võib keerulisemate/ressursimahukamate teisenduste puhul tõsta

oluliselt jõudlust. Mitmetuumalised protsessorid on saanud juba üsna tavaliseks ning uus loodav tarkvara peaks suutma neid efektiivselt ära kasutada.

3.3.3 Kasutajaliides ja visuaalsed komponendid

Kasutajaliidese loomisel tuleb arvestada, et see peab olema veebipõhine, st toimima enamkasutatavates brauserites, ning olema dünaamiliselt muudetav. Viimane omadus on vajalik terminikirje muutmisliideses, kuna väljade struktuur igas terminikirjes võib olla erinev ja peab olema redigeerimisel muudetav – välju peab saama kirje redigeerimise käigus lisada ja eemaldada.

Uute väljade defineerimise võimalus koos terminiandmete sisestamisega (sisestatakse välja nimi ja väärtus koos) ning automaatne nimistute loomine sisestatud väärtuste põhjal (kui välja tüübiks on valitud nimistu, siis välja väärtuse sisestamisel pakutakse juba varem sellise nimega väljale sisestatud väärtusi).

3.3.3.1 Vaated

Kasutajaliidese vaated:

- Kirjevaade
- Sõnastikuvaade
- Otsingutulemused

3.3.3.2 Andmeväljad

Andmeväljadel esitatakse andmeelemente, mis vastavad andmekategooriatele.

Lihtandmetüüpide kuvamiseks ja muutmiseks, nagu tekst, loendist valik, kuupäev ja kellaaeg, leidub komponente pea kõikides kasutajaliidese raamistikes. Keerulisem on leida aga vahendeid hierarhilise dünaamilise vormi kuvamiseks. Selleks võiks kasutada XForms mootorit.

Veel on vajalik visuaalne komponent mõistesüsteemide graafiliseks kujutamiseks.

3.4 Kasutatavad välised teenused

Väliste teenuste all on mõeldud funktsioone, mis jäävad antud töö uurimisteemade alt välja, kuid mille järele vajadus terminihaldussüsteemis kindlasti tekib. Selliseid funktsioone võib käsitleda terminihaldussüsteemi suhtes kui infrastruktuuri

komponente. Ei oma erilist tähtsust, kas neid funktsioone kasutatakse XML-teenustena või sisalduvad nad näiteks THS-is endas programmiteekidena. Kuna lähtume teenuspõhisest arhitektuurist, nimetame neid teenusteks. Selliste teenustena võib välja tuua keeletehnoloogilised vahendid:

- Eesti keele speller, mida saaks kasutada veebiteenusena igasugune rakendus.
- Lemmatiseerija. Vajalik selleks, et otsida terminibaasist näiteks definitsioonides leiduvaid termineid, mis ei pruugi olla algvormis või otsida termineid, mis ei ole otsingu sisendisse antud algvormis.
- Morfoloogiline analüsaator. Vajalik nii spelleris kui lemmatiseerijas, lisaks ka terminituvastuses ja terminite ekstraheerimisel. Kahte viimast rakendust eesti keele jaoks veel ei ole olemas. Hetkel olemasolev EKI-s loodud morfoloogiline analüsaator vajab uuendamist [Willemson07].

3.5 Tarkvara platvormi ja komponentide valik

Tarkvara komponendid tuleks valida lähtudes eelnevalt välja toodud vajadustest.

Keele ja haridusasutustel probleemiks finantsvahendite puudus. Tuleb otsida avatud või tasuta lahendusi. Süsteemi haldamine ei tohiks nõuda väga palju inimtööjõuressurssi.

Ideaalis tuleks valida tarkvarakomponendid, mis ei maksa, on avatud, millele on leida piisavalt dokumentatsiooni ja tuge ja on niivõrd levinud, et ei nõua eriväljaõppega spetsialiste.

Eelnevalt tutvustatud XML-andmebaasisüsteemidest sobivaimaks võib pidada DB2-te.

Vahevara platvorm tuleb valida silmaspidades järgnevaid vajadusi: arenduslihtsus, arendajate kättesaadavus tööjõuturul, teabe kättesaadavus, XML-i töötluse võimalused, veebiteenuste tugi, lõimede tugi, sõltumatus operatsioonisüsteemist, tehnoloogia perspektiivikus jne.

DB2-l on andmebaasiga ühendumiseks võimalik kasutada nii Java kui ka .NET API-sid, lisaks on andmebaasile juurdepääs ka veebiteenuseid kasutades. Veebiteenuseid kasutades oleks vahevara lihtsalt teatud funktsionaalsusega veebiteenuste vahendaja.

Kasutajaliides tuleks realiseerida kasutades olemasolevaid AJAX teeke.

4 Praktiline osa

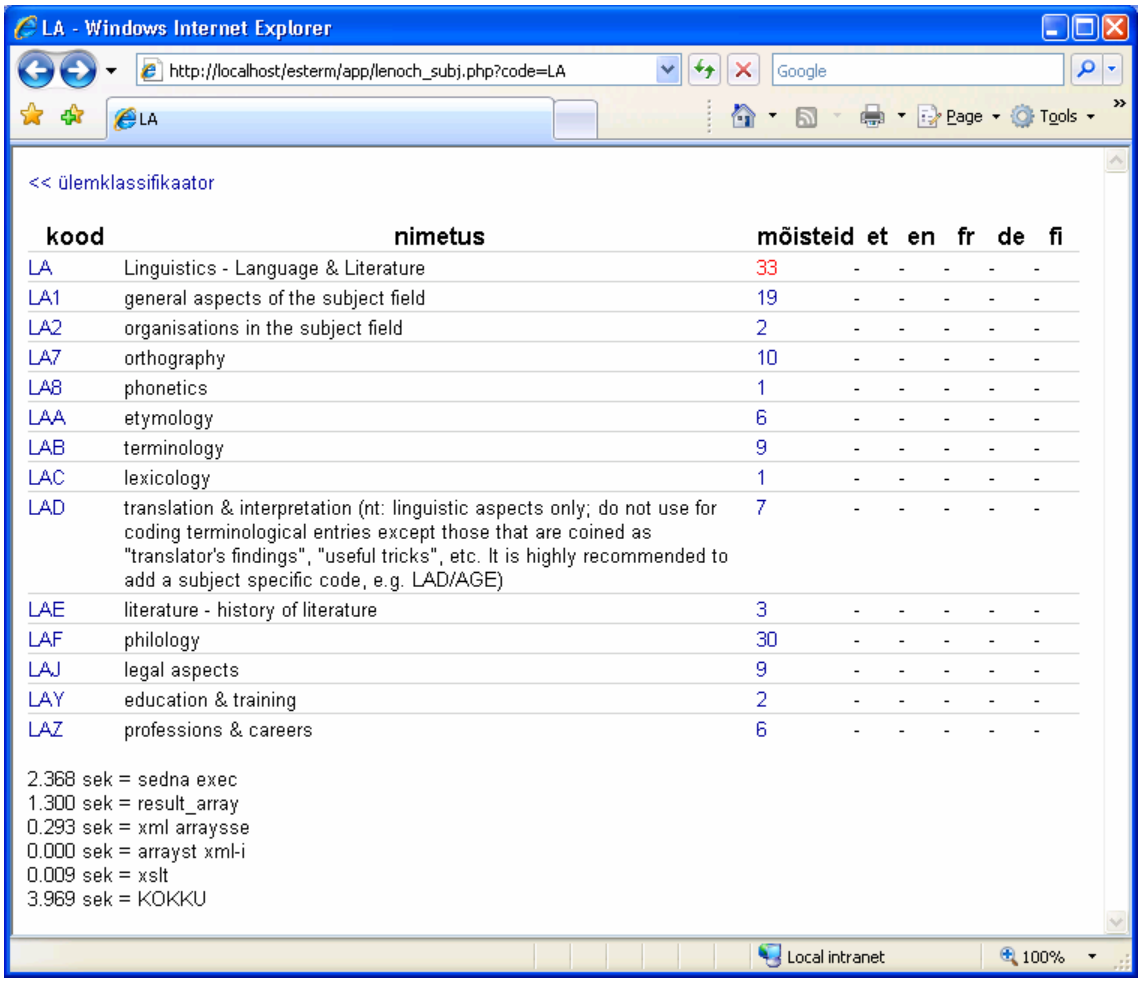
4.1 Prototüübid

4.1.1 Terminibaasi avalik päringuliides

ETKs kasutatava terminihaldussüsteemi Multiterm veebipõhine komponent Multiterm Online töötab vaid brauseriga MS Internet Explorer, mis ei võimalda teda kasutada näiteks Linux platvormil. See on oluline puudus üldkasutatava terminibaasi avalikustamisel. Lisaks on Multiterm Online üheaegsete kasutajate arv piiratud serveri litsentside arvuga. Samas on Multiterm Online veebiliides ohtra Javascripti kasutamise tõttu liialt ressursinõudlik ning ebatöökindel. Kuna selliste piirangute tõttu ilmnes palju probleeme, otsustas käesoleva töö autor luua Multiterm Serveri baasil uue veebiliidese, küsides selleks luba ka Multitermi tootjafirmalt. Viimane oli nõus, kuna oli teadlik probleemidest. Uus nn kohandatud veebiliides (vt Joonis 6) on realiseeritud kasutades Multiterm Serveri API-t ja ASP veebiskripte. Veebibrauserilt saabuv päring edastatakse Multiterm Serveri API-le, millelt saadav vastus on XML-fragment. XSLT-teisendusega muudetakse XML HTML-iks ja edastatakse brauserile.

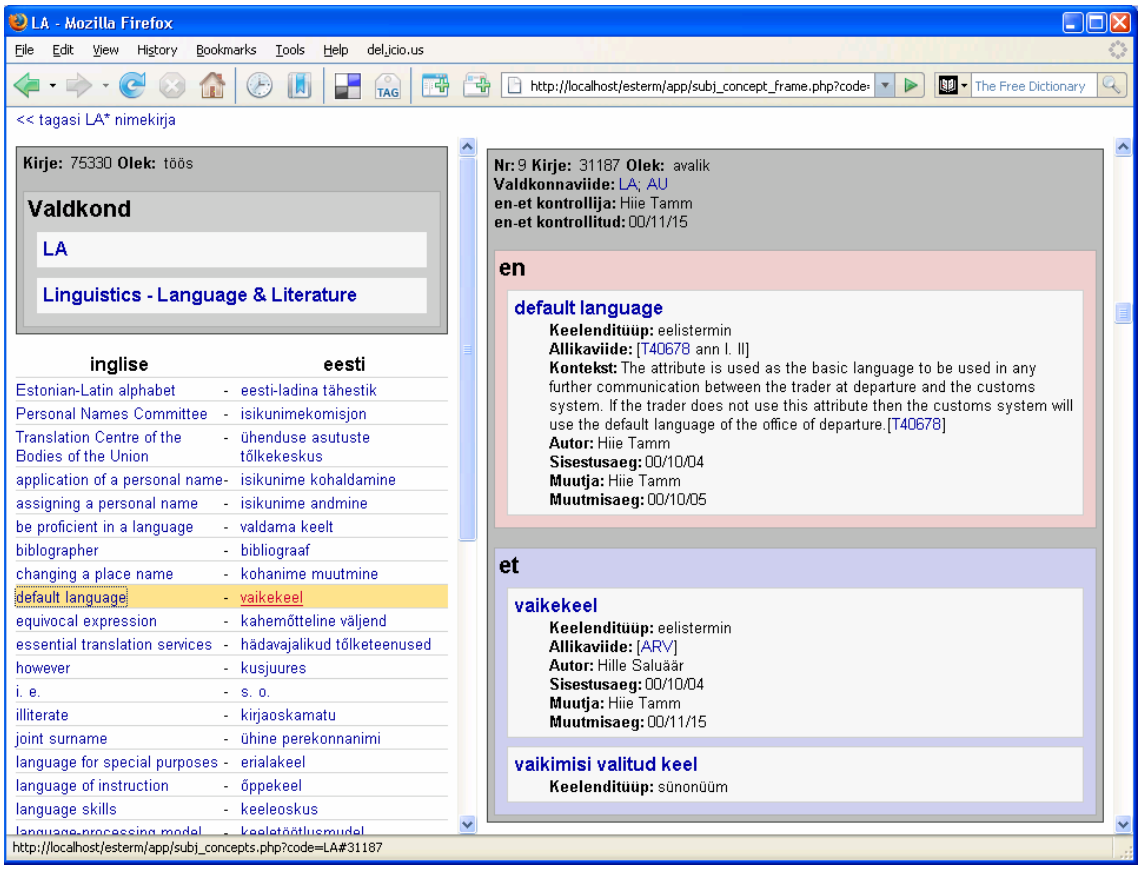
4.1.2 Terminibaasi valdkonnapõhine sirvimine – *proof of concept*

XML-andmebaasi testimiseks reaalses rakenduses on loodud terminibaasi Esterm jaoks valdkondade statistika ja sirvimise veebipõhine rakendus. Realiseeritud on rakendus kasutades Sedna algupärast XML-andmebaasi, PHP skriptingkeelt ja XQuery päringuid. Rakendus loendab ja filtreerib terminibaasi XML-põhistest terminiartiklitest valdkonna märgendeid ja võimaldab kogu terminibaasi termineid grupeerida valdkondade kaupa, kuvades seal juures kirjete arvu igas valdkonnas. Valdkonnad on omakorda jaotatud ülemvaldkondadeks ja alamvaldkondadeks (vt Joonis 10).



Joonis 10. Terminibaasi Esterm valdkonnapõhise sirvimise rakendus. Kassifikaatori vaade.

XML-andmebaasi testimiseks reaalses rakenduses on loodud terminibaasi Esterm jaoks



Joonis 11. Terminibaasi Esterm valdkonnapõhise sirvimise rakendus. Valdkonna terminite vaade.

Kokkuvõte

Bakalaureusetöö eesmärgiks oli uurida olemasolevaid terminihaldussüsteeme ja anda ülevaade tarkvara funktsionaalsustest, mida terminoloogiatöös vajatakse, eriti ETK vajaduste kontekstis, samuti anda ülevaade XML-andmebaasisüsteemidest. Töös uuriti XML-andmebaaside kasutamisvõimalusi terminihaldussüsteemides nii teoreetilisest kui ka praktilisest vaatepunktist lähtudes.

Töös pakuti välja tehnoloogiaid, mida kasutada ja produkte, mida kohandada, et luua vajalik uus terminihaldussüsteem ning eesti terminoloogia edendamiseks vajalik avalik terminihalduskeskkond.

Töö praktilise osana realiseeriti prototüübid XML-põhistest terminibaasi liidestest, millest üks kasutab olemasoleva kommertsiaalse tarkvara API-t ja teine vabavaralist XML-andmebaasi.

Summary

Using an XML Database in a Terminology Management System

The main focus of this thesis is a theme which falls within the shared area of terminology and information technology. The information technology needs of the sphere of terminology are mainly related to the collection, storing, processing and forwarding of information. For the realisation of all these needs, different terminology management systems (TMS) have been created. A database plays a key role in a terminology management system because it is the characteristics of the database which define the functions which the TMS can offer for users. In most cases, the existing TMS-s have been constructed on the basis of relational databases. However, a new field of database systems, XML databases, is being developed increasingly.

The aim of this thesis is to study the potential advantages of using an XML database particularly in the case of TMS-s.

The thesis introduces the main principles of the structure of a term base. It is established that the general data model of a term entry is hierarchical in its structure and, at the same time, the structure should be easily modifiable because adding and modifying the data fields of the entries is a daily necessity. The XML language, being hierarchical and structured by nature and, at the same time, dynamic and flexible, is very well suited for presenting such data entry. In order to allow for more effective saving and querying of the XML, the thesis provides a study on the use of an XML database as a TMS back end.

The thesis is divided into four chapters.

The first chapter provides an overview of terminology work in general, the existing terminology management systems and the basic principles in the designing of term databases (term bases).

In the second chapter, the existing XML database systems are introduced and their essential characteristics and usability for the purposes of term bases is studied.

The third chapter provides an overview of a universal open terminology management system which is being designed at present and describes the basic requirements therefor and the structure thereof.

The fourth chapter describes the practical use of XML databases with the example of provisional prototypes.

Kasutatud kirjandus

[ISO 1087-1] EVS-ISO 1087-1:2002 TERMINOLOOGIATÖÖ. SÕNASTIK Osa 1: Teooria ja rakendus

[ISO 1087-2] EVS-ISO 1087-1:2002 TERMINOLOOGIATÖÖ. SÕNASTIK Osa 2: Arvutirakendused

[Erelt07] Tiiu Erelt. (2007). Terminiõpetus.

[EOS] Tiiu Erelt, Arvi Tavast. (2003). Eesti oskuskeelekorralduse seisund. Eesti Keele Sihtasutus, Tallinn

[Tavast02] Arvi Tavast. (2002). Terminibaasi koostamise põhimõtted. http://www.imprimaatur.ee/artiklid/kiirylev.html

[Lauk05] Siiri Lauk. (2005). Oskuskeelekorraldus: terminibaasid Eestis ja mujal maailmas. Referaat. http://evkk.tlu.ee/pdfs/term_ref.pdf

[Seewald-Heeg06] Uta Seewald-Heeg. (2006). Terminology Exchange without Loss? (2006). Lk 5 – 18. http://www.ldv-forum.org/2006_Heft1/LDV-Forum1.2006.pdf

[Loopmann07] Andres Loopmann. Sõnastike haldussüsteem EELex. http://dspace.utlib.ee/dspace/bitstream/10062/2931/1/loopmann_andres.pdf

[Kaeeli04] Tiit Kaeeli. XML dokumentide andmebaasisüsteemid. http://www.egeen.ee/u/vilo/edu/Students/Tiit_Kaeeli/Tiit_Kaeeli_Dipl.pdf

[Einama06] Kaido Einama. Keskkonna sõbralik hübriidmootoriga andmebaas. Arvutimaailm dets 2006

[e-teatmik] H. Vallaste, E-teatmik: IT ja sidetehnika seletav sõnaraamat. http://www.vallaste.ee/ (viimati vaadatud 31.12.2007).

[Wenzel] Annemette Wenzel. (2006). i-Term – more than a termbase. Terminfo (2/2006) 14-17

[Buysschaert] Joost Buysschaert. The Development of a MeSH-based Biomedical Termbase at Hogeschool Gent. http://estime.spim.jussieu.fr/~pz/lrec2006/Buysschaert.pdf

[Kaalep07] H. J. Kaalep. http://keeletehnoloogia.cs.ut.ee/konverents/slaidid/kaalep-keeleveeb.pdf

[ecolore] Lokaliseerimisprotsessi ja tööriistade ülevaade. Tööriistad: Terminihaldus. http://ecolore.leeds.ac.uk/xml/materials/overview/tools/terminology_management.xml?lang=et

[Kuub02] Tarmo Kuub bakalaureusetöö "Eesti seaduste struktureerimine XML tehnoloogia abil", Tallinn 2002

[Toova04] Tanel Toova proseminaritöö "Tõlketarkvarad. Andmekorje lisamoodul Trados 6.0-le", Tallinn 2004

[Alegria] I. Alegria, X. Arregi, X. Artola, M. Astiz. A DICTIONARY CONTENT MANAGEMENT SYSTEM

http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1150789608/publikoak/dictionary_content.pdf

[visthes] VisualThesaurus.com

[Wordnet] http://wordnet.princeton.edu/

[Esterm] Estermi veebiliides. http://mt.legaltext.ee/esterm/

[Sutrop07_1] Urmas Sutrop, Kaur Männiko, Andres Loopmann. „Ühtne Terminoloogiasõnaraamatute Koostamise Keskkond". Projektitaotlus.

[HTM05] Haridus- ja Teadusministeerium. Riiklik programm „Eesti keele keeletehnoloogiline tugi (2006-2010)". Tartu 2005 http://www.riigikantselei.ee/failid/Keeletehnoloogia_riiklik_programm.doc

[Termbases] http://www.termbases.eu/

[i-Term] http://www.i-term.dk/

[ETK06_1] Probleemid ja vajadused ESTERMis. Eesti Terminoloogia Keskus. Elektrooniline dokument sisekasutuseks.

[ETK06_2] Probleemid ja vajadused MILITERMis. Eesti Terminoloogia Keskus. Elektrooniline dokument sisekasutuseks.

[Sutrop07_2] Eestikeelse terminoloogia toetamise riiklik programm (2008-2012). Urmas Sutrop. 2007. Asutustevaheline elektrooniline dokument.

[ETER] Eesti Terminoloogia Ühingu koduleht. http://www.eter.ee/ (viimati vaadatud 31.12.2007)

[ETER07] Oskuskeele korralduse andmebaas. http://www.eter.ee/andmebaas/ (viimati vaadatud 31.12.2007)

[Willemson07] Jan Willemson, Jaak Pruulmann-Vengerfeldt. Reeglipõhine keeletarkvara. (2007). EKKTT konverents, Ettekande slaidid. http://keeletehnoloogia.cs.ut.ee/konverents/slaidid/pruulmann.pdf

[rpbourret] http://www.rpbourret.com/xml/XMLDatabaseProds.htm

[Cuong] Nguyen Viet Cuong

http://swing.felk.cvut.cz/index.php?option=com_docman&task=doc_view&gid=5&Itemid=62