Käyttöohje

Halaan-ryhmä

Helsinki 18.12.2006

Ohjelmistotuotantoprojekti

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Kurssi

581260 Ohjelmistotuotantoprojekti (6 ov)

Projektiryhmä

Paula Kemppi
Liisa Oikarinen
Ville Palkosaari
Maria Rinta-Opas
Jaakko Sorri
Marjaana Välisalo

Asiakas

Terttu Nevalainen

Johtoryhmä

Ilari Moilanen
Kimmo Simola

Kotisivu

`http://www.cs.helsinki.fi/group/ohtu/halaan`

Versiohistoria

Versio	Päiväys	Tehdyt muutokset
0.1	23.11.2006	Sisällysluettelon luonnos
0.11	24.11.2006	Rakenteen muokkaus, korpustietojen lisäyksestä
0.12	26.11.2006	Sisällön päivitys
0.2	29.11.2006	Sisällön päivitys
0.21	13.12.2006	Lukuja 5 ja 7.2
0.22	14.12.2006	Luku 3
0.23	15.12.2006	Luku 7.1 ja liite 1
1.0	17.12.2006	Sisällön muokkausta ja oikolukua

Sisältö

1 Johdanto

Tämä dokumentti on CEECer-hakukoneen käyttöohje. CEECer on VARIENG-tutkimusyksikön käyttöön suunniteltu www-käyttöliittymällä varustettu hakukone laajalle annotoidulle tekstiaineistolle. Aineisto muodostuu vanhoista englanninkielisistä kirjeistä ja näihin liittyvästä metadatasta. Hakukone mahdollistaa sosiolingvistiset haut, eli kirjeiden ja niihin liittyvien tietojen hakemisen tietokannasta esimerkiksi kirjoittajan sukupuolen, asuinpaikan ja kirjoitusajankohdan perusteella. Järjestelmä tarjoaa myös muun muassa mahdollisuuden muokata hakutuloksen esitystä ja tallentaa haulla rajattu osaaineisto tai vain siihen liittyvä metadata. Hakutuloksesta voidaan luoda myös pylväsdiagrammi kuvaamaan tuloksen jakaumaa eri parametreilla. Korpukseen voidaan lisätä uutta aineistoa tiedostosta ja olemassaolevaa voidaan päivittää sekä tiedoston kautta että www-käyttöliittymästä. Järjestelmän käyttö edellyttää kirjautumista ja myös käyttäjätietojen hallinnointi on osa järjestelmää.

2 Sanasto

Admin Käyttäjäryhmä, jolla on kaikki oikeudet; ylläpitäjä; pääkäyttäjä; ylläpitotyökalut sisältävä sivu.

Apache Ant Työväline automatisoimaan toistuvia toimintosekvenssejä.

Boolean- tai koodityyppinen arvo Arvo, jonka arvoalue on rajattu (boolean arvoihin tai tiettyihin koodeihin).

J2EE-sovelluspalvelin *Java Enterprise Editionin* mukainen palvelinohjelmisto.

Java5 SDK Java-ohjelmointikielen version 1.5 *Standard Development Kit*.

Luokitteleva parametri Graafisen esityksen vaaka-akselille sijoittuva parametri.

Toolbar Valikko, jonka kautta navigointi järjestelmän pääsivujen sisällä on mahdollista.

URL *Uniform Resource Locator* Yksikäsitteinen osoite verkkoresurssiin, yleiskielessä www-osoite.

3 Järjestelmän käyttöönotto

Loppukäyttäjä käyttää järjestelmää selaimen kautta. Näin ollen loppukäyttäjältä ei vaadita varsinaisia asennustoimenpiteitä. Loppukäyttäjän käytössä on oltava nykyaikainen selain sekä tiedossa URL, josta hakukone löytyy. Tässä luvussa keskitytään palvelimen asetusten säätämiseen.

3.1 Ohjelmistovaatimukset

Koneesta, johon järjestelmä asennetaan, tulee löytyä seuraavat ohjelmistot:

- J2EE-sovelluspalvelin
- Tietokanta
- Java5 SDK
- Apache Ant

Sovelluspalvelinta ja tietokantaa tarvitaan järjestelmän käyttämiseksi. Java5 SDK ja Ant ovat välttämättömiä järjestelmän uudeelleenkääntämiseen lähdekoodista ja asennuspaketin luomiseen.

3.2 Asennuksen valmistelut

Järjestelmän käyttöönotto vaatii muutamia muutoksia *ceec.properties*-tiedostoon ceec-hakemistoon. Tiedostoon on päivitettävä tiedot tietokanta-ajurista ja -palvelimesta sekä käyttäjätunnus ja salasana tietokantaan eli seuraavat kohdat:

- dbDriver
- dbServer
- dbUser
- dbPassword

Lisäksi samasta tiedostosta on päivitettävä *adminMail*, joka on siis pääkäyttäjän sähköpostiosoite. Tämä osoite tulee hakukoneen kirjautumissivulle *Register here* -linkkiin mailto-osoitteeksi. Sähköpostiosoitetta lukevan oletetaan siis huolehtivan käyttäjien lisäämisestä järjestelmään. Myös luokkaan ceec.container.StringContainer on seuraavien muuttujien arvoksi laitettava asennusympäristöä vastaavat arvot:

- ROOT-muuttuja
- PROPERTYFILENAME-muuttuja

Jotta järjestelmä voidaan asentaa, käytössä on oltava ceec.war-tiedosto. War-tiedosto on pakattu tiedosto, joka sisältää kaikki järjestelmään kuuluvat komponentit. Tarvittaessa war-tiedosto voidaan koota Ant-ohjelman avulla:

1. Siirrytään ceec-hakemistoon, josta löytyy build.xml-tiedosto.
2. Komento `ant war` tuottaa war-tiedoston.

3.3 Tietokannan luominen

Järjestelmä vaatii toimiakseen tietokannan, toteutuksessa käytössä on ollut Oracle. Tietokannan luomiseen tarvittavat komennot löytyvät tiedostosta *create_ceec_db.sql*. Kun kyseinen tiedosto on suoritettu käytössä olevan tietokannan komentotulkissa, tietokannan taulut on luotu.

Taulujen luomisen jälkeen suoritetaan tiedosto *insert_ceec_db.sql*, joka alustaa tietokannan: tietokantaan lisätään käyttäjäryhmätiedot sekä luodaan yksi pääkäyttäjä. Pääkäyttäjän tunnuksilla (käyttäjätunnus admin, salasana ADMIN) pääsee kirjautumaan sisään järjestelmään web-käyttöliittymän kautta. Web-käyttöliittymän kautta järjestelmään pystyy lisäämään uusia käyttäjiä ja pääkäyttäjän salasana on myös syytä vaihtaa.

3.4 Asennus ja asetusten säätäminen

Järjestelmä asennetaan kopioimalla war -tiedosto J2EE-sovelluspalvelimen webapps-hakemistoon. Kaikki sovelluspalvelimet eivät osaa ottaa war-tiedostoa automaattisesti käyttöön, vaan vaativat esimerkiksi uudelleenkäynnistyksen. Asia on syytä tarkistaa käytettävän sovelluspalvelimen dokumentaatiosta. Jos lähdekoodia muokkaa tässä vaiheessa (esim. StringContainerin muuttujien arvoja), koodin kääntäminen onnistuu ceec-hakemistossa komentamalla *ant*.

Lisäksi Tomcat-ympäristössä on tarpeen luoda ceec.xml -niminen tiedosto webapps-hakemistoon täydentämään Tomcat-konfiguraatiota (tai määrittää sovelluspolut /home/[username]/tomcat/conf/server.xml -tiedostoon). Tiedostossa ideana on liittää ulospäin näkyvä URL sisäiseen polkunimeen. Sen sisältö voi näyttää esimerkiksi seuraavalta:

```
<Context
        path="/tomcat/[username]/ceec"
        docBase="/home/[username]/tomcat/webapps/ceec/"
        debug="0"
        reloadable="true"
        crossContext="true"
        override="true"/>
```

4 Käytön aloitus

4.1 Rekisteröityminen

Järjestelmän käyttäjäksi rekisteröidytään lähettämällä sähköpostia *Login*-sivulta löytyvään ylläpidon osoitteeseen (*Register here* -linkki). Tämän jälkeen järjestelmän ylläpitäjä luo järjestelmään uudet tunnukset ja lähettää hakijalle käyttäjätunnuksen ja salasanan.

4.2 Sisäänkirjautuminen (Login)

Järjestelmään kirjaudutaan sisään syöttämällä *Login*-sivulle käyttäjätunnus ja salasana niille kuuluviin kenttiin.

4.3 Valikkorivi (Toolbar)

Järjestelmän eri osien välillä voidaan navigoida valitsemalla valikosta haluttu toiminto. Tarjolla olevat toiminnot vaihtelevat sen mukaan, millaiset käyttöoikeudet käyttäjällä on järjestelmään.

1. **Search Page** Tältä sivulta on mahdollista suorittaa hakuja. Sivu aukeaa ensimmäiseksi sisäänkirjautumisen yhteydessä.
2. **Import Data** Tältä sivulta voidaan järjestelmään lisätä uutta tietoa, tai jos lisättävä tieto on jo olemassa, päivittää sitä. Näkyy vain Admin- ja CEEC-käyttäjäryhmille.
3. **User Administration** Tämän linkin takaa löytyvät ylläpitäjän työkalut käyttäjänhallintaan. Näkyy vain Admin-käyttäjäryhmälle.
4. **Help (Finnish)** Linkki vie hakukoneen suomenkieliseen käyttöohjeeseen.
5. **Logout** Tätä klikkaamalla kirjaudutaan ulos järjestelmästä.

5 Hakeminen

5.1 Uusi haku

Sisäänkirjautumisen jälkeen käyttäjä ohjataan hakusivulle, josta uuden haun luominen onnistuu täyttämällä hakukaavake ja valitsemalla *Search*. Jos käyttäjä on jo tehnyt hakuja, mutta haluaa tehdä uuden haun edellisen muokkaamisen sijaan, se onnistuu klikkaamalla valikosta *Search page* -linkkiä, joka tyhjentää lomakkeen valituista tiedoista.

5.2 Tallennettujen hakujen käyttö

Aiemmin tallennetun haun käyttäminen on mahdollista valitsemalla hakusivun yläreunassa olevasta *Saved searches* -pudotusvalikosta haluttu haku ja painamalla *Do Query* -painiketta, jolloin kaavake täyttyy tallennetun haun parametrien mukaan. Tämän jälkeen voidaan normaalisti valita lomakkeen lopusta *Search*, jolloin haku suoritetaan ladatuilla parametreillä. Tallennetun haun lataamisen jälkeen hakulomakkeen valintoja voidaan vielä muuttaa ennen haun tekemistä. Haku on mahdollista poistaa valitsemalla se valikosta ja painamalla *Delete Query* -painiketta. Hakujen tallennus kuvataan tulossivun yhteydessä luvussa 6.1.

5.3 Hakuparametrit

Kaikki hakuparametrit ovat rajaavia.

5.3.1 Search For

Search For -valikosta valitaan, mitä haetaan. Letter hakee kirjeitä, Person ihmisiä, Collection kokoelmia, Sender kirjeiden kirjoittajia ja Recipient kirjeiden vastaanottajia.

Henkilöitä (Person) hakiessa ei oteta lainkaan huomioon Recipient Specific Settings -osiossa olevia mahdollisia valintoja, vaan haetaan Person Settings -parametreja vastaavat henkilöt. Jos on valittu myös kirjeiden ja/tai kokoelmien parametreja, haetaan ne henkilöt, jotka ovat parametreja vastaavissa kirjeissä kirjoittajina tai vastaanottajina.

Lähettäjiä, vastaanottajia ja kirjeitä hakiessa haetaan hakuehtoja vastaavat kirjeet ja hakutulokseen tulee valinnasta riippuen joko näiden kirjeiden kirjoittajat, vastaanottajat tai itse kirjeet.

Kokoelmia hakiessa, jos on valittu pelkästään kokoelmaparametreja, haetaan näitä parametreja vastaavat kokoelmat. Jos taas on valittu myös henkilö- tai kirjeparametreja, haetaan parametreja vastaavat kirjeet ja tulokseen tulevat ne kokoelmat, joissa nämä kirjeet sijaitsevat.

5.3.2 Letter Text Version

Letter Text Version -valikossa valitaan, mikä tekstiversio kirjeistä täytyy olla olemassa. Tämä on siis yksi kirjeen hakuparametreista: haetaan ne kirjeet, joista löytyy valittu versio. Version valinta määrää myös sen, mikä versio kirjeteksteistä voidaan exportata. Jos haetaan henkilöitä niin, että vain henkilöparametreja on valittu, tai kokoelmia niin, että vain kokoelmaparametreja on valittu, ei tekstiversion valintaa oteta huomioon mitenkään.

6 Tulossivu

6.1 Haun tallennus (Add Query)

Haun tallennus on mahdollista tulossivun yläosasta löytyvän *Add Query* -painikkeen avulla. Haulle on hyvä ensin antaa kuvaava nimi Description-kenttään. Tämä toiminto tallentaa viimeksi tehdyn haun myöhempää käyttöä varten. Tallennetut haut ovat käyttäjäkohtaisia. Hakujen tallennusmahdollisuutta ei ole Guest- ja GuestLite-käyttäjäryhmillä.

6.2 Tuloslistan katselu ja muokkaus

Tulossivulla näytetään listamuotoinen esitys hakuun täsmänneistä kokoelmista, henkilöistä tai kirjeistä. Tuloksen rivejä voidaan järjestää tietyn parametrin mukaan klikkaamalla

kyseisen parametrin nimeä. Sarakkeita on mahdollista piilottaa parametrin nimen vieressä olevasta [*]-linkistä. Sarake ei piiloudu kokonaan, vaan kyseinen linkki jää näkyviin. Vietäessä hiiri linkin päälle tulee näkyviin piilotetun sarakkeen nimi ja klikkaamalla sitä sarakkeen saa uudelleen näkyviin.

6.3 Tuloslistan tallennus (Export Data)

Tuloslistan voi tallentaa *Export List* -painikkeesta. Lista on tab-erotellussa muodossa.

6.4 Osakorpuksen tallennus (Export Text)

Jos tuloslista koostuu kirjeistä, listan alapuolelta löytyy myös *Export Text* -painike, jota painamalla voi ladata haun rajaaman osakorpuksen itselleen. Tekstiversio on sama kuin mihin haku on hakulomakkeessa määritelty kohdistuvaksi (Letter Text Version -parametri hakulomakkeessa).

6.5 Graafinen esitys (Create Graph)

Hakutuloksesta on mahdollista luoda graafinen esitys valitsemalla tulossivulta *Create Graph*. Tämä toiminto kuvataan luvussa 7.

6.6 Korpustietojen päivitys (Update Data)

Mikäli käyttäjä kuuluu Admin- tai CEEC-käyttäryhmään, tuloslistan alaosasta löytyy myös *Update Data* -painike, joka mahdollistaa korpustietojen muokkauksen ja poiston. Tämä toiminnallisuus kuvataan luvussa 8.2.

7 Graafinen esitys

7.1 Luonti

Hakutuloksesta on mahdollista luoda graafinen esitys valitsemalla tulossivulta *Create Graph*. Painike on tarjolla, mikäli hakutulos koostuu henkilöistä tai kirjeistä. Napin painalluksen jälkeen graafinen esitys eli pylväsdiagrammi aukeaa uuteen ikkunaan. Pylvään korkeus määräytyy henkilöhauissa henkilöiden lukumäärän perusteella ja kirjehauissa aina kirjeiden sisältämien sanojen lukumäärän perusteella. Oletusarvoisesti graafin luokittelevana parametrina on aika ja yhden luokan koko on viidennes tarkasteltavasta ajanjaksosta. Kirjeiden kohdalla aika tarkoittaa kirjeen kirjoitusvuotta ja henkilöiden kohdalla syntymävuotta.

Ajanjakso, jolta kuvaaja esitetään, on oletusarvoisesti se aikaväli, jolta kirjeitä tai henkilöitä hakutuloksessa esiintyy. Mikäli kirjeen kirjoitusvuosi tai henkilön elinvuodet eivät ole tiedossa, nämä eivät näy kuvaajassa. Henkilö- ja kirjehauissa kirjoitus-, syntymä- tai kuolinvuosien rajoittaminen hakulomakkeessa määrää myös kuvajaan alku- ja/tai loppuajankohdan. Lähettäjä- tai vastaanottajahaussa vuosia ei huomioida. Vuodet eivät välity kuvaajaan, mikäli tuloslistaa järjestää jonkin parametrin mukaan ennen kuvaajan piirtämistä.

7.2 Muokkaus

Pylvään kuvaaman ajanjakson pituutta voidaan säätää joko määrittämällä pylväiden määrä tai suoraan yhden pylvään kuvaaman ajanjakson pituus. Nämä kentät ja *Modify*-nappi löytyvät kuvaajan alapuolelta silloin, kun kuvaaja on muodostettu vuosiluvun pohjalta. Epäsuorempi muokkaus on mahdollista hakulomakkeen avulla. Jos haetaan esimerkiksi Basire-kokoelman kaikki kirjeet kirjoitusvuosia rajoittamatta ja piirretään kuvaaja, aikajaksoksi tulee 1651-1666. Oletuksena piirrettävät viisi pylvästä tarkoittavat kolmea vuotta pylvästä kohden lukuunottamasta viimeistä, johon tulee neljä vuotta. On hyvä huomata, että viimeinen pylväs poikkeaa vuosien suhteen aina, kun jako ei mene tasan (tarkan vuosimäärän näkee pylvään alle tulostuvasta tekstistä). Jos tehdään sama haku, mutta määrätään kirjoitusvuosiksi esimerkiksi 1650-1670, ei rajata hakutulosta, mutta voidaan määrittää kuvaajan alku vuoteen 1650 ja loppu vuoteen 1670.

Luokittelevaa parametria on myös mahdollista vaihtaa. Parametreiksi tarjotaan pudotusvalikossa sellaisia parametrejä, jotka saavat boolean- tai koodityyppisiä arvoja poikkeuksena kuolinvuosi henkilökuvaajan yhteydessä. Uuden kuvaajan voi piirtää *New Graph* -painikkeesta. Kuvaajaan tulee pylväät kustakin koodista sekä yksi pylväs niitä kirjeitä tai henkilöitä varten, joilla kyseinen parametri on tyhjänä. Tämän pylvään alla on tekstinä kysymysmerkki.

7.3 Tallennus

Kuvan tallentaminen onnistuu viemällä hiiri kuvan päälle, klikkaamalla hiiren oikeaa painiketta ja valitsemalla esiin tulevasta valikosta *Save Image As...*. Kuva on png-muodossa.

8 Korpustietojen hallinnointi

8.1 Korpustietojen lisäys (Import Data)

Korpustietojen lisääminen tietokantaan edellyttää, että käyttäjä kuuluu Admin- tai CEEC-ryhmiin. Tällöin valikosta löytyy *Import Data* -linkki. Se vie sivulle, joka tarjoaa lomakkeen, josta voi ladata päivitykset sisältävän tiedoston. Tiedoston lataus ja tietojen tallennus kuvataan luvuissa 8.1.5-8.1.6. Korpustietojen lisäys tietokantaan tapahtuu siis aina

tiedostosta. Tietokantaan voidaan lisätä kahdenlaista tietoa: metadataa (kokoelmia, henkilöitä tai kirjeitä) ja itse kirjetekstejä. Metadata on tallennettava muotoon, jossa sarakkeet on eroteltu sarkaimella eli tabilla. Järjestelmään tuotavan aineiston tyypin tunnistus perustuu tiedostonimeen.

8.1.1 Metadatatiedostojen muodosta ja sisällöstä

Tiedoston tyypin tunnistus perustuu tiedoston nimeen (ks. luku 8.1.3) ja parametrien tunnistus perustuu sarakkeiden nimiin. Nimien on vastattava sovittuja ja sarakenimet luetellaan liitteessä 1. Kirjainkoolla ei kuitenkaan ole merkitystä. Tietokannassa pakollisiksi määritellyt kentät vaaditaan eli näiden nimet eivät voi puuttua. Ylimääräisiä tietokantaan kuulumattomia parametreja järjestelmä ei huoli. Tämä on lähinnä siksi, että esimerkiksi kirjoitusvirheen sattuessa ei-pakollisen parametrin nimeen, ei kyseinen sarake jää käyttäjän huomaamatta tallentumatta kantaan. Sarakkeiden keskinäisellä järjestyksellä ei ole väliä. On myös hyvä huomata, että taulukko-ohjelmassa ns. piilotetut sarakkeet tulevat mukaan tallennuksessa tab-eroteltuun muotoon ja voivat aiheuttaa ikäviä yllätyksiä.

Kentistä pakollisiksi määritellyt eivät voi olla tyhjiä, muut voivat. Tiedostossa mahdollisesti olevat tyhjät rivit jätetään huomiotta. Pituus- ja sisältörajoitteet kuvataan liitteessä 1. Niin sanottujen boolean-tyyppisten kenttien arvon tulee olla *Y* tai *N*. On tärkeää, että kentissä ei ole tekstin tai lukujen seassa välilyöntejä tai tabeja. Mikäli näitä esiintyy, tiedosto tallentuu väärin ja sen jäsentäminen oikein on mahdotonta. Tällaisen tiedoston syöttäminen järjestelmään aiheuttaa luultavasti kasan virheilmoituksia, jotka johtuvat kenttien sisältörajoitusten rikkoumisesta. Järjestelmä yrittää etsiä mahdollisesti katkenneet rivit ja virheilmoituksesta löytyy näiden rivinumerot. Ylimääräisten välilyöntien poistaminen poistaa luultavasti myös ison osan sisältöä koskevista virheilmoituksista. Mahdolliset välilyönnit (\n) ja tabit (\t) on helpointa poistaa ennen tiedoston tallennusta taulukko-ohjelman *Edit*-valikon *Find & Replace* -toiminnolla tai vastaavalla.

Tiedostoa jäsentävä parseri olettaa, että parametrin arvoa mahdollisesti ympäröivät lainausmerkit ovat Excelin tai OpenOfficen aikaansaannoksia. Itse tekstidatassa mahdollisesti olevat lainausmerkit nämä ohjelmat tuplaavat tallennuksessa. Toisin sanoen parseri poistaa yksinkertaiset lainausmerkit ja yksinkertaistaa kaksinkertaiset. Tämän ei pitäisi näkyä käyttäjälle mitenkään, mutta käsittelystä on hyvä olla tietoinen siltä varalta, että esimerkiksi intoutuu rakentamaan tab-eroteltuja tiedostoja käsin.

8.1.2 Kirjetekstitiedostojen muodosta ja sisällöstä

Yhdessä tiedostossa voi olla useita kirjetekstejä kuitenkin niin, että kaikki tekstit ovat samaa versiotyyppiä eli joko plain (P), tagged (T) tai parsed (S). Kirjeen version on käytävä ilmi tiedostonimestä (ks. luku 8.1.3). Kirjeet talletetaan kantaan yksittäin ja niiden erottelu perustuu kirjeiden alussa oleviin tageihin, joista voidaan poimia sen kirjeen tunniste, johon kirje kuuluu.

Plain-version kirje alkaa, kun riviltä löytyy `<L ` -alkuinen tagi, esim. `<L BASIRE_001>`. Tagged-version kirje alkaa `<L_`-alkuisen tagin sisältävällä rivillä, esim.

<Q_BAS_D_1651_FN_FBASIRE>_CODE <L_BASIRE_001>_CODE
Parsed-versiossa kirje tunnistetaan alkavaksi, kun törmätään <Q_-alkuiseen tagiin ja seuraavalta riviltä löytyy <L_-alkuinen tagi, esim.
((CODE <Q_BAS_D_1651_FN_FBASIRE>))
((CODE <L_BASIRE_001>))
Kaikkien tagien pitää tietysti loppua >-merkkiin. Yksi kirje tulkitaan loppuneeksi, kun kohdataan seuraavan alku tai tiedoston loppu.

Kirjeen sisältä mahdollisia tyhjiä riviä ei poisteta, mutta lopuista nämä poistetaan. Lisäksi poistetaan sovitusti kaikki mahdolliset <SAMPLE (plain) tai <S_SAMPLE (muut versiot) -alkuiset tagit ylimääräisinä ja turhina.

8.1.3 Tiedoston tallennus vaadittavaan muotoon ja tiedostonimet

Excel-tiedosto tallennetaan valitsemalla *File*-valikosta *Save as*. Tiedostomuodoksi valitaan *Save as type*-valikosta *Text (tab delimited)*, jolloin sarakkeet erotetaan toisistaan sarkaimella. Open Officessa tiedostomuodoksi pitää valita *Text CSV (.csv)*, erotinmerkiksi (Field delimiter) *{Tab}* ja merkkijonoja ympäröimään (Text delimiter) tavallinen lainausmerkki. Metadatatiedostojen nimien on loputtava collection.txt, person.txt tai letter.txt, tosin tiedostopääte voi olla myös csv. Nimen alussa voi siis olla lisätietoa, esim. uuttadataa_collection.txt on käypä tiedostonimi kokoelmadatalle.

Kirjetekstitiedostot ovat tavallisia tekstitiedostoja. Tiedostojen nimistä on sovittu, että plain-versiossa ei ole tiedostopäätettä, tagged-versio päättyy .pos ja parsed-versio .psd. Näiden tiedostojen nimille ei ole asetettu muita rajoitteita. Esimerkiksi BASIRE, basire.pos ja basire.psd toimivat. Kirjainkoolla ei ole merkitystä missään tiedostonimissä, mutta ääkkösiä ja välilyöntejä ei tule käyttää.

8.1.4 Tiedostojen tuontijärjestys

Korpustietojen keskinäiset suhteet asettavat vaatimuksia uuden aineiston tuotijärjestykselle. Ensin järjestelmään tuodaan joko kokoelmien tai henkilöiden tiedot. Kirjeisiin liittyvien kokoelmien ja henkilöiden on oltava tietokannassa ennen kuin kirjeitä voidaan lisätä. Kirjetekstit lisätään tietokantaan kirjeiden jälkeen.

8.1.5 Tiedoston lataus

Mikäli käyttäjällä on riittävät oikeudet, valikosta löytyy *Import Data* -linkki. Se vie sivulle, joka tarjoaa lomakkeen, josta voi ladata päivitykset sisältävän tiedoston. Import Corpus Data -lomakkeen kenttään kirjoitetaan ladattavan tiedoston nimi tiedostopolkuineen. Kirjoittamista helpompi tapa on etsiä tiedosto *Browse...* -nappia painamalla (tässä voi lukea myös *Selaa...*). Tämä avaa erillisen ikkunan, josta voi tutkia paikallisen koneen hakemistoja. Haluttu tiedosto voidaan valita tuplaklikkaamalla tai valitsemalla tiedosto ja painamalla *Avaa* -painiketta tai vastaavaa. Tämä siirtää tiedoston nimen mainitun lomakkeen kenttään, jonka jälkeen painetaan *Import*-nappia. Jos tiedoston nimessä on jo-

tain vikaa tai tiedoston jäsentämisessä tapahtuu virheitä, sivulle tulostuu virheilmoituksia, joiden perusteella tiedostoon voi tehdä muutoksia.

8.1.6 Uuden tiedon tuonti ja olemassa olevien tietojen korvaus

Mikäli tiedosto pystytään jäsentämään, näytetään listaus datasta ennen kuin se viedään tietokantaan. Jos tiedostosta tuotu data on uutta eli tunnisteilla ei löydy mitään tietokannasta, listan otsikko on *New data*. Listassa on jokaisella rivillä *Modify* ja *Remove* -painikkeet. Ensinmainitusta pääsee lomakenäkymään, jossa tietoja on vielä mahdollista muokata ja *OK*-napista muutokset siirtyvät listaan. Jälkimmäinen poistaa rivin listasta. Kirjetekstejä ei ole mahdollista muokata.

Tiedostosta voi uuden datan sijaan tai lisäksi tuoda myös tietoja, jotka ovat jo tietokannassa. Tällaiset tiedostosta saadut tiedot näytetään *Data to Update* -listassa samalla sivulla. Tarjolla on täysin samat muokkaus- ja poistotoiminnot. Molemmat listat, joista toinen on tyhjä, jos data on pelkästään uutta tai vanhaa, viedään tietokantaan listojen alapuolelta löytyvästä *Save*-napista. Uuden datan tapauksessa tiedot lisätään tietokantaan, olemassaolevan tapauksessa ne kirjoitetaan tietokannassa olevien päälle.

8.2 Korpustietojen muokkaus ja poisto (Update Data)

Korpustietojen muokkaus tapahtuu tavallisen haun avulla. Ensin käyttäjä hakee haluamansa datan hakulomakkeella. Mikäli käyttäjätunnuksella on riittävät oikeudet eli se kuuluu Admin- tai CEEC-käyttäjäryhmään, näkyy hakutuloksen näyttävän sivun alalaidassa *Update Data* -nappula, jota klikkaamalla käyttäjä siirtyy tietojenmuokkaussivulle. Muokkaussivulla näytetään hakutulos listana, jossa on ne tiedot, joita käyttäjä voi muokata. Muokattavia eivät ole yksilöivät tunnukset eivätkä laskettavat yhteenvetotiedot. Jokaisella rivillä on *Modify* ja *Delete* -nappulat. Tietojen muokkaus tapahtuu klikkaamalla *Modify* -nappulaa, minkä jälkeen käyttäjälle näytetään valitun datarivin tiedot lomakkeella, jolla tietojen muokkaus on mahdollista. Käyttäjän antamat tiedot tallentuvat tietokantaan painamalla lomakkeen alareunassa olevaa *Save*-nappulaa.

Korpustietojen poisto tapahtuu samalta tietojenmuokkaussivulta klikkaamalla *Delete*-nappulaa. Nappulan painaminen poistaa datarivin kokonaan tietokannasta. Poistettaessa kokoelmia myös kokoelmaan kuuluvat kirjeet poistetaan sisältöineen. Poistettaessa henkilöitä myös henkilön kirjoittamat ja vastaanottamat kirjeet poistetaan. Poistettaessa kirjeitä myös kirjeiden tekstit poistetaan.

9 Käyttäjätietojen hallinnointi

9.1 Käyttäjän lisäys

Uusi käyttäjä lisätään valitsemalla valikosta *User Administration*, joka vie käyttäjänhallintaan. Sivu näyttää listan järjestelmän käyttäjistä. Uusi käyttäjä lisätään valitsemalla *Add*

User. Lomakkeella on kentät etu- ja sukunimelle, sähköpostiosoitteelle ja puhelinnumerolle (kaikki maksimissaan 50 merkkiä), käyttäjärymälle (pudotusvalikko) sekä käyttäjätunnukselle (enintään 8 merkkiä) ja salasanalle (enintään 10 merkkiä). Näistä pakollisia ovat käyttäjätunnus, etu- ja sukunimi, salasana ja käyttäjäryhmä.

9.2 Käyttäjäryhmät

Käyttäjäryhmät luodaan tietokannan luonnin yhteydessä. Käyttäjäryhmä määrittää käyttäjän käyttöoikeudet järjestelmään. Ryhmien kuvaukset ovat seuraavanlaiset:

- **Admin** Kaikki oikeudet, ei aineistorajoitetta.
- **CEEC** Kuten Admin, mutta ei oikeutta käyttäjätietojen hallinnointiin.
- **Researcher** Kuten CEEC, mutta ei oikeutta korpustietokannan päivityksiin.
- **ResearcherLite** Kuten Researcher, mutta aineistorajoite voimassa.
- **Guest** Kuten ResearcherLite, mutta ei oikeutta tallentaa hakuparametreja tietokantaan.
- **GuestLite** Kuten Guest, mutta ei oikeutta osakorpuksen ja metadatan tallentamiseen.

Aineistorajoite viittaa siihen, että osalla aineistosta on rajoitettu lukuoikeus. Kirjeiden, joiden copyright-parametrin arvo on *N*, lukuoikeus on vain käyttäjäryhmillä Admin, CEEC ja Researcher. Muiden käyttäjäryhmien haut rajautuvat automaattisesti kirjeaineistoon, jossa copyright-parametrin arvo on *Y* eli näillä aineistorajoite on voimassa.

9.3 Käyttäjätietojen muokkaus

Käyttäjätietojen muokkaus on mahdollista valitsemalla valikkorivistä *User Administration*. Tällä sivulla on lista kaikista järjestelmän käyttäjistä. Jokaisen käyttäjän tietojen oikealla puolella on *modify*-nappi. Tätä painamalla avautuu ikkuna, jossa on mahdollista muuttaa käyttäjän tietoja. Muutokset saatetaan voimaan painamalla *modify*-nappia. Huomioon tulee ottaa, että viimeisen Admin-ryhmään kuuluvan käyttäjän ryhmää ei pysty muuttamaan. Sivulta voidaan poistua ilman muutosta painamalla CANCEL-linkkiä lomakkeen yläpuolella.

9.4 Käyttäjän poisto

Käyttäjien poistaminen on mahdollista valitsemalla valikkorivistä *User Administration*. Tällä sivulla on lista kaikista järjestelmän käyttäjistä. Jokaisen käyttäjän tietojen oikealla puolella on *delete*-nappi. Sitä painamalla avautuu dialogi, jossa poisto varmistetaan. Valitsemalla *OK* käyttäjä poistetaan.

10 Uloskirjautuminen

Järjestelmästä on mahdollista kirjautua ulos valitsemalla valikkoriviltä *Logout*.

Liite 1. Parametrien nimet, maksimipituudet ja koodit

Alla olevissa taulukoissa luetellaan tietokantaan tulevien parametrien nimet, niiden maksimipituudet ja sisällölliset rajoitteet. Tyypeistä varchar ja char viittaavat mihin tahansa merkkeihin, number tarkoittaa numeerista arvoa. Jos tyyppi on char(1) eikä koodeja ole annettu, viittaa tämä boolean-tyyppiseen kenttään, joka voi saada arvoksi *Y* tai *N*. Alleviivaus nimessä tarkoittaa, että kyseinen parametri on tunniste eli sen pitää olla yksilöllinen. Pituuksien yhteydessä *not null*-määreet kertovat, että kenttä ei voi olla tyhjä.

Kirje		
LetterID	varchar(15), not null	tunniste
Collection	=Collection:Name, not null	kokoelma
Sender	=Person:PersonCode, not null	kirjoittaja
SenderRank	varchar(3)	kirj. sos. status (1)
SenderStatus	varchar(255)	kirj. sos. status
MultiSenders	char(1)	useita kirjoittajia
Recipient	=Person:PersonCode, not null	vastaanottaja
RecRank	varchar(3)	vo. sos. status (1)
RecStatus	varchar(255)	vo. sos. status
MultiRec	char(1)	useita vastaanottajia
Year	number(4)	kirjoitusvuosi
YearUncertain	char(1)	vuosi epävarma
WordCount	number(5)	sanamäärä
RelCode	varchar(2)	kirj-vo -suhde (2)
Relationship	varchar(255)	kirj-vo -suhde
Place	varchar(255)	kirjoituspaikka
Authenticity	varchar(3)	kirjeen autenttisuus (3)
LetterDate	varchar(255)	aika tarkemmin vapaamuotoisesti
AddressFormula	char(1)	aloituskaava
ClosingFormula	char(1)	lopetuskaava
ContentType	varchar(255)	sisältötyyppi
LetterNotes	varchar(1000)	muuta kirjeestä
CorrespondentNotes	varchar(1000)	muuta henkilöistä
LetterNumber	varchar(50)	numero editiossa
PageNumber	varchar(50)	sivunumero editiossa
Source	varchar(500)	lähde
Copyright	char(1)	julkaisuoikeus
Complete	char(1)	tiedot valmiit
Updated	date	viimeisin päivitys
NewBoolean1	char(1)	-
NewBoolean2	char(1)	-
NewText1	varchar(50)	-
NewText2	varchar(255)	-
NewNumber	number(5)	-

Koodit:

1. SenderRank / RecRank: samat kuin Person-taulun Rank ja FatherRank -kentillä
2. RelCode: FN (family nuclear), FO (family other), FS (family servant), TC (close friend), T (other)
3. Authenticity: A (holograph), B (holograph; writer's social background partly unknown), C (later copy), D (uncertain authenticity; copy & writer's social bg partly unknown), E (modernized), S (scribal/secretarial) tai 2 näistä, esim. AC, CA (+?)

Henkilö		
PersonCode	varchar(20), not null	yksilökoodi
Sex	char(1)	sukupuoli (1)
Region	char(1)	asuinalue (2)
County	varchar(5)	asuinkunta (3)
SocMob	char(1)	sos. liikkuvuus (4)
LastName	varchar(255)	sukunimi
FirstName	varchar(255)	etunimi
TitleOccupation	varchar(255)	titteli
YBirth	number(4)	syntymävuosi
YDeath	number(4)	kuolinvuosi
Lifespan	varchar(255)	elinaika epävarmuusilmauksin
Rank	varchar(3)	korkein sos. status (5)
FatherRank	varchar(3)	isän sos. status (5)
Father	varchar(1000)	isän tiedot
PBirth	char(1)	syntymäpaikka (2)
MigCode	varchar(3)	muuttohistoria (6)
Migration	varchar(1500)	muuttohistoria
EduCode	varchar(4)	koulutus (7)
Education	varchar(1500)	koulutus
Career	varchar(1500)	ura
Religion	char(1)	uskonto (8)
DNB	varchar(50)	linkki ODNB Onlineen
Notes	varchar(1500)	lisätietoa
SentLettcont	varchar(10)	kirj. kirjeiden sis.tyypit (9)
RecLettcont	varchar(10)	v-o. kirjeiden sis.tyypit (9)
Complete	char(1)	tiedot valmiit
Updated	date (yyyy-mm-dd)	viimeisin päivitys
NewBoolean1	char(1)	-
NewBoolean2	char(1)	-
NewText1	varchar(50)	-
NewText2	varchar(255)	-
NewNumber	number(5)	-

Koodit:

1. Sex: F (female), M (male)
2. Region / PBirth: N (north), F (east anglia), H (home counties), L (london), C (court), O (other), A (abroad)
3. County: BDF, BKM, BRK, CAM, CHS, CON, CRT, CUL, DBY, DEV, DOR, DUR, ESS, GLS, HAM, HEF, HRT, HUN, KEN, LAN, LEI, LIN, LND, MDX, NBL, NFK, NTH, NTT, OXF, RUT, SAL, SFK, SOM, SRY, SSX, STS, WAR, WES, WIL, WOR, YKS, ERY, WRY, NRY, CHI, IOM, IOW, ABR (+A, C, F, H, L, N, O)
4. SocMob: U (up), D (down), N (none)
5. Rank / FatherRank: R (royalty), N (nobility), GU (gentry upper), GL (gentry lower), G (gentry), P (professional), CU (clergy upper), CL (clergy lower), M (merchant), O (other), (+?)
6. MigCode: Y (yes), YL (yes: london), YA (yes: abroad), YLA (yes: london & abroad)
7. EduCode: A (apprenticed), E (elementary), H (higher), HC (higher: cambridge), HI (higher: inns of court), HO (higher: oxford), PC (private/self: classical), PN (private/self: non-classical), S (secondary), HF (higher: foreign), (+ C, O, I, F) (+?)
8. Religion: P (protestant), A (anglican), C (catholic), X (unknown)
9. SentLettcont / RecLettcont: M (mixed), B (business), P (private), N (news), O (official), W (other), L (love), D (duty), T (travel), F (family) - mikä tahansa yhdistelmä näistä *(kentän sisältöä ei tarkisteta)*

Kokoelma		
Name	varchar(20), not null	nimi
Filename	varchar(8)	tiedostonimi
FromYear	number(4)	vuodesta
ToYear	number(4)	vuoteen