Antti Sievänen

Ammattinimikkeiden automaattinen lajittelu SQL-tietokannassa

Metropolia Ammattikorkeakoulu
Insinööri (AMK)
Tietoteknikka
Insinöörityö
11.4.2011

Metropolia

Tiivistelmä

Tekijä Otsikko Sivumäärä Aika	Antti Sievänen Ammattinimikkeiden automaattinen lajittelu SQL-tietokannassa 35 sivua 11.4.2011
Tutkinto	Insinööri (AMK)
Koulutusohjelma	Tietotekniikka
Suuntautumisvaihtoehto	Ohjelmistotekniikka
Ohjaajat	Kehityspäällikkö Matti Jokela Lehtori Vesa Ollikainen

Bisnode Finland Oy toimii Väestörekisterikeskuksen toimittamien henkilötietojen jälleenmyyjänä ja saa osana henkilötietoja henkilön itsensä ilmoittaman ammattinimikkeen. Työn tarkoitus oli toteuttaa Bisnode Finland Oy:lle automatisoitu järjestelmä, joka lajittelisi kirjoitusmuodoiltaan vaihtelevat ja erimuotoiset ammattinimikkeet niitä vastaaviin virallisiin ammattiluokituksiin. Tämän avulla ammattinimikkeitä voitaisiin entistä paremmin hyödyntää esimerkiksi asiakkaalle poimittavia kohderyhmiä poimittaessa.

Tuloksena kehitettiin helposti ylläpidettävä ja tarpeiden mukaan muokattavissa oleva järjestelmä, jonka avulla entistä suurempi osa henkilöiden ilmoittamista ammattinimikkeistä saadaan kohdentumaan virallisiin ammattinimikkeisiin ja niiden kautta ammattiluokkiin. Tämä lisää Bisnode Finland Oy:n kilpailukykyä tarjoamalla helpomman keinon rajata kohderyhmiä käyttäen hyväksi henkilön ammattinimikettä.

Avainsanat	SQL, automaattinen, merkkijonovertailu

Metropolia

Abstract

Author Title Number of Pages Date	Antti Sievänen Automatic sorting of profession titles in SQL-database 35 pages 11 April 2011
Degree	Bachelor of Engineering
Degree Programme	Information Technology
Specialisation option	Software Engineering
Instructors	Matti Jokela, Development Manager Vesa Ollikainen, Senior Lecturer

Bisnode Finland Oy acts as a reseller of personal data provided by Population Register Centre and as a part of the personal data Bisnode gets a person's profession title which has been provided by the person himself. The purpose of the project was to implement a system for Bisnode Finland Oy that would automatically sort mistyped titles and titles spelled in various different ways into official profession title classes. After that the profession title classes could more easily be made use of when compiling target groups to customers.

As a result a system was created which is easy to maintain, can be modified according to needs and is able to match a higher percentage of profession titles, that people have entered into the population register, to their official titles and hence their official profession classes. This improves Bisnode Finland Oy's competitiveness by providing an easier way to confine target groups using person's profession title.

Keywords	SQL, automatic, string comparison

Sisällys

Käsiteluettelo

Funktio	SQL:n avulla toteutettu skripti, joka suorittaa valmiiksi tallennetun käsittelyn ajossa annettujen parametrien perusteella ja palauttaa aina jonkin tuloksen.
Jobi	SQL:n avulla toteutettu vaiheista koostuva komentojoukko, joka voi sisältää tallennettuja funktioita ja proseduureja, SSIS-paketteja tai SQL-komentoja.
Proseduuri	SQL:n avulla toteutettu skripti, joka suorittaa valmiiksi tallennetun toiminnallisuuden. Proseduuriin voidaan antaa parametreja, ja se voi palauttaa tietoa.
SQL	*Structured Query Language*. IBM:n kehittämä standardoitu kyselykieli, jolla voidaan tehdä hakuja, muutoksia ja lisäyksiä relaatiotietokantaan.
SSIS-paketti	*SQL Server Integration Services* -paketti on erityisen kehittimen avulla luotu kokonaisuus, joka suorittaa sen sisään määritellyt tehtävät SQL-ympäristössä.
Taulu	Sarakkeista ja riveistä koostuva taulukkomainen tietokantaobjekti, johon voidaan tallentaa kyselyn avulla saatuja tuloksia.
VRK	Väestörekisterikeskus on viranomaistaho, joka yhdessä maistraattien kanssa pitää yllä väestötietojärjestelmää.
VTJ	Väestötietojärjestelmä on valtakunnallinen atk-rekisteri, joka sisältää perustiedot Suomen kansalaisista ja Suomessa vakinaisesti asuvista ulkomaalaisista.

1 Johdanto

Työni aiheena on ammattinimikkeiden automaattinen lajittelu SQL-tietokannassa. Tavoitteena on tarjota Bisnode Finland Oy:n käyttöön järjestelmä, joka tuotantoon siirrettäessä käsittelee viikoittain kaikkien väestötietojärjestelmän täysi-ikäisten, markkinointikelpoisten henkilöiden erilaiset ammattinimikkeet. Tämän jälkeen järjestelmä lajittelee ne automaattisesti valmiisiin, Tilastokeskuksen koostamiin ammattinimikkeisiin ja sitä kautta eri ammattiluokkiin.

Projekti alkoi Bisnode Finland Oy:n aloitteesta, kun tutkittiin mahdollisuuksia kohdentaa erilaiset ammattinimikkeiden kirjoitusmuodot valmiiseen ammattinimikkeistöön. Vanhaa, tähän asti käytössä ollutta, ammattinimikeluetteloa ei nähty järkeväksi käyttää, sillä vuosien mittaan uusia ammatteja on kehityksen myötä tullut lisää, ja nimikkeet ovat muuttuneet. Täten päätettiin käyttää uutta ammattinimikelistausta, joka hankittiin Tilastokeskukselta.

Tilastokeskuksen ammattinimikkeisiin on laadittu hierarkia, jonka avulla yksittäinen ammattinimike saa neljästä viiteen merkkiä pitkän numerokoodin, joista jokainen koodin numero tarkentaa ammattia. Otetaan esimerkiksi henkilö, jonka ammattinimike on ”autotarvikemyyjä” ja joka saa näin ammattinimikelistauksesta koodin 52203. Ensimmäinen numero (5) kertoo henkilön kuuluvan ”palvelu-, myynti- ja hoitotyöntekijöihin”, seuraava numero (2) tarkentaa henkilön kuuluvat ryhmään ”mallit, myyjät ja tuote-esittelijät”. Tämän jälkeinen numero (2) tiputtaa vielä mallit pois ko. ryhmästä. Kaksi viimeistä numeroa (03) tarkentaa henkilön kuuluvan lopulta ryhmään ”erikoismyyjät”. Nimensä mukaisesti samaan ryhmään kuuluu tämän lisäksi lukuisia erikoisalojen myyjiä. Koska koodin pituus on viisi numeroa, on ammattinimike niin ikään tason viisi eli tarkimman tason nimike Tilastokeskuksen ammattinimikelistauksessa.

Tämän hierarkian avulla kohderyhmää rajatessa voidaan edellä mainitun esimerkin tapauksessa poimia joko pelkät autotarvikemyyjät koodilla 52203. Jos kohderyhmää kuitenkin halutaan laajentaa hakemalla kaikki myyjät ja tuote-esittelijät, voidaan käyttää hakuehtona, jossa koodin kolmen ensimmäisen merkin pitää olla 522. Samalla logiikalla käyttämällä hakuehtoa, jossa koodin ensimmäisen merkin pitää olla 5, saadaan

haettua kaikki palvelu-, myynti- ja hoitotyötä tekevät henkilöt. Tämä mahdollistaa kohderyhmien tietojen poiminnan huomattavasti helpommin ja kattavammin kuin ennen.

Esittelen aluksi hieman taustoja sekä tietojentoimittajasta, Väestörekisterikeskuksesta, että työnantajastani Bisnode Finland Oy:stä. Tämän jälkeen paneudutaan tarkemmin tarpeeseen toteuttaa tämä työ sekä itse ongelmaan ja etuihin, joita tämä järjestelmä toimiessaan tuo yritykselle.

Tästä edetään työn suunnitteluun sekä toteutukseen eri vaiheineen. Lopuksi perehdytään työn tuloksiin sekä niiden vertailuun ja niistä ilmeneviin parannusideoihin, jotka jäävät toteutettaviksi tämän työn ulkopuolella.

Henkilötietolakiin on Suomessa kirjattu, että yksityishenkilöiden osoitetietoja voidaan viranomaisen luvalla luovuttaa esimerkiksi suoramarkkinointiin, mikäli henkilö itse ei ole erikseen kieltänyt tietojensa luovuttamista (1). Tämä mahdollistaa liiketoiminnan, jossa yritys Väestörekisterikeskuksen luvalla poimii väestötietojärjestelmästä erilaisia kohderyhmiä annettujen kriteerien perusteella ja myy niitä yksityisille asiakasyrityksille.

Bisnode Finland Oy toimii muun muassa tällaisena jälleenmyyjänä, joka poimii väestötietojärjestelmästä erilaisia kohderyhmiä rajaamalla niitä yleisimmin muun muassa henkilön iän, kielen, sukupuolen ja asuinpaikkakunnan perusteella.

Kohderyhmien rajaukseen on käytettävissä myös henkilön itsensä ilmoittama ammattinimike, joka voidaan syöttää väestötietojärjestelmään joko muuton yhteydessä täytettävässä muuttoilmoituksessa, ilmoittamalla uusi ammattinimike maistraattiin tai ilmoittamalla uusi ammattinimike VRK:n Tarkasta tietosi! – Internet-palvelun kautta (2). Eniten ilmoituksia ammattinimikkeestä tulee muuttoilmoituksen kautta, joka on lain mukaan tehtävä aina, kun muuttaa pysyvästi asunnosta toiseen tai kun tilapäinen oleskelu toisessa osoitteessa kestää yli kolme kuukautta (3), mutta muuttoilmoituksen yhteydessä ammattinimikkeen ilmoittaminen on vapaaehtoista.

Nämä kolmea eri kautta saadut tiedot henkilöiden ammattinimikkeistä liitetään väestötietojärjestelmään osaksi henkilöiden tietoja ja henkilötietojen jälleenmyyjä saa myös tämän tiedon käyttöönsä yhdeksi poimintaperusteeksi.

2 Ongelman kuvaus

Ammattinimikkeet luetaan Väestörekisterikeskuksessa sisään joko maistraatin toimittamasta sähköisestä aineistosta tai skannataan käsin täytetystä muuttoilmoituksesta ja luetaan samalla tekstiksi. Ammattinimikkeitä ei ole järjestelmään syötettäessä sidottu mihinkään valmiisiin, standardeihin nimikkeisiin, vaan henkilö saa itse syöttää ammattinimikkeen vapaamuotoiseen tekstikenttään. Henkilön kirjoittamaa ammattinimikettä ei tarkisteta mitenkään, vaan ammattinimike rekisteröityy VRK:lle sellaisenaan, mikäli skannauksen yhteydessä teksti pystytään lukemaan sisään ilman ongelmia. Tämän takia väestötietojärjestelmästä löytyvien eri ammattinimikkeiden kirjoitusmuotoja on todella suuri määrä, eivätkä kaikki syötetyt ammattinimikkeet välttämättä edes ole varsinaisia ammattinimikkeitä, vaan mukana on esimerkiksi erilaisia koulutukseen liittyviä titteleitä.

Edellä mainittujen seikkojen lisäksi englanniksi tai ruotsiksi kirjoitetut ammattinimikkeet sekä kirjoitusvirheet kasvattavat kirjoitusmuotojen määrää entisestään. Tällä hetkellä erilaisia kirjoitusmuotoja väestötietojärjestelmästä löytyy 166 389 kappaletta ja määrä kasvaa jatkuvasti. Tämä tekee ammattinimikkeiden käytöstä yhtenä poimintaehtona lähes mahdotonta, sillä yhdellä, oikeinkirjoitetulla ammattinimikkeellä, saadaan osumaan vain vaihteleva osa henkilöistä, jotka todellisuudessa kuuluisivat ko. ammattiin.

Otetaan esimerkkinä ammattiryhmä lääkärit. Tilastokeskuksen ammattinimikeluettelosta löytyy 65 eri ammattinimikettä haettaessa ammattinimikkeitä, joista löytyy merkkijono ”lääkäri”. Väestötietojärjestelmästä samalla haulla löytyy 638 eri ammattinimikettä, ja jos lisäksi sallitaan hakuehdoksi vaihtoehtoisesti nimikkeet, joista löytyy merkkijono ”lääk.” tai jotka loppuvat merkkijonoon ”lääk”, eli yleiset lyhenteet lääkärille, määrä nousee 880 kappaleeseen. Edelleen hakemalla kaikki henkilöt, joiden ammattinimike on joku edellä mainituista 880 kappaleesta, saadaan tuloksena 22 040 henkilöä. Täten jos kohderyhmää halutaan rajata hakemalla kaikki lääkärit, saadaan ilman minkäänlaista luokitusta hakutuloksena 12 294 henkilöä eli noin 56 % mahdollisesta kohderyhmästä. Todellisuudessa lääkärien lukumäärä väestötietojärjestelmässä on vielä suurempi, sillä tässä esimerkissä ei ole otettu huomioon esimerkiksi kirjoitusvirheitä, joita sisältävät nimikkeet jäävät puhtaalla merkkijonovertailulla löytymättä. Suomen lääkäriliiton mukaan 1.1.2011 Suomessa asuvia lääkäreitä oli 22 821 kappaletta (4).

Erityisen vaikeaa on käsitellä ammattinimikkeitä, joita on dramaattisesti lyhennetty (esimerkiksi ”mv” = maanviljelijä), joiden virallinen kirjoitusasu eroaa suuresti henkilön itse ilmoittamasta (esimerkiksi ”toimistoduunari” = toimistotyöntekijä) tai joissa kenttään on syötetty useampi ammattinimike (esimerkiksi autonkuljettaja/ravintolatyöntekijä). Ammattinimikkeiden lajittelun suurin ongelma yleisesti onkin se, että vapaamuotoinen tekstikenttä, johon henkilöt ammattinsa kirjaavat, sallii käytännössä minkä tahansa syötetyn tekstin. Täten ammattinimikkeeksi on voitu syöttää esimerkiksi täysin sattumanvarainen merkkijono, ja koska koneellisesti virheellisiä merkkijonoja on lähes mahdotonta erikseen tunnistaa, on järjestelmän yritettävä lajitella ne johonkin valmiiksi annetusta ammattinimikelistasta.

Ammattinimikkeiden tulkinnassa tarpeellista on myös epäselvien tapausten etsiminen esimerkiksi Internetistä, sillä monesti vastaan tulee selkeästi kirjoitettuja ammattinimikkeitä, joiden merkitys ei kuitenkaan ole heti selvä. Sama ongelma saattaa ilmetä joidenkin lyhenteiden kohdalla.

Vaihtoehtona on tietysti ammattinimikkeen lajittelematta jättäminen, joskaan se ei tietysti ole toivottava vaihtoehto. Alusta asti oli kuitenkin selvää, että epäselvät tapaukset kannattaa jättää ennemmin lajittelematta kuin lajitella väärin, sillä lajittelematta jääneiden ammattinimikkeiden läpikäyminen olisi huomattavasti helpompaa kuin virheellisten synonyymiparien etsiminen ja korjaaminen.

Ammattinimikkeiden lajitteluun on vuonna 1997 tehty käsin lista erilaisista kirjoitusmuodoista, johon suoraan vertaamalla saadaan nimikkeeseen kiinni ammattiluokitus. Tätä listaa ei ole ylläpidetty enää pitkään aikaan, joten kaikki uudet kirjoitusmuodot ammattinimikkeistä sekä kokonaan uudet ammatit jäävät ilman ammattiluokitusta ja tippuvat näin pois poiminnoista, jotka rajataan käyttäen hyväksi ammattiluokkaa. Näitä vanhoja synonyymeja läpi käydessäni löysin myös muutamia tapauksia, joissa synonyymi ja ammattinimike eivät vastanneet toisiaan eli synonyymi on ilmeisesti lisätty väärin inhimillisen virheen takia.

Tämän työn tarkoituksena oli kehittää automaattinen järjestelmä, joka yrittää sijoittaa syötetyn ammattinimikkeen johonkin Tilastokeskuksen muodostamista virallisista ammattinimikkeistä, jonka jälkeen yksi ammattinimike, joka koostuu useasta eri kirjoitus-

muodosta, voidaan poimia väestötietojärjestelmästä yhdellä viralliseen ammattinimikkeeseen liitetyllä koodilla.

Tässä ammattinimikelistauksessa erilaisia suomenkielisiä ammattinimikkeitä on 8 764 kappaletta sekä lisäksi 7 412 ruotsinkielistä ammattinimikettä. Myös englanninkielisiä ammattinimikkeitäkin esiintyy Suomen väestössä jonkin verran, mutta valmista ammattinimikelistausta englanniksi ei ollut saatavilla, joten englanninkielisten synonyymien muodostus jäi käsin tehtäväksi. Kaikki ammattiluokat on jaettu lisäksi ylempään luokitukseen, joka kertoo kyseisen ammattiluokan sosioekonomisen aseman. Myös tätä luokitusta voidaan käyttää hyväksi kohderyhmien tietojen poiminnassa.

Ammattinimikettä yhtenä poimintaehtona käytettäessä on kuitenkin otettava huomioon, ettei se ole parhaita tapoja rajata kohderyhmää. Mitään takuuta ammattinimikkeen oikeellisuudesta ei ole, sillä henkilö voi syöttää kyseiseen kenttään mitä tahansa. Toinen ammattinimikkeen luotettavuuteen vaikuttava asia on se, että jos henkilö ei muuta, pysyy ammattinimike todennäköisesti samana vuodesta toiseen, sillä yleensä tieto ammattinimikkeestä saadaan juuri henkilön ilmoittamasta muuttoilmoituksesta. Hyvin harva vaivautuu muuten ilmoittamaan uudesta ammatistaan, sillä se ei ole mitenkään pakollista eikä siitä hyödy mitenkään.

3 Aikaisemmat ratkaisut

Aiheen perimmäinen ongelma, merkkijonojen vertailu keskenään, on hyvin yleinen ongelma. Ongelma esiintyy yleisesti tilanteissa, joissa merkkijonolle yritetään löytää vastinetta ehdokkaiden joukosta. Tiedostamattamme tähän törmää lähes jokapäiväisessä elämässä esimerkiksi yrittäessämme saada selvää epäselvästä käsialasta tai vaikka etsiessämme osittain kuultua osoitetta kartalta. Näitä edellä mainittuja esimerkkejä yhdistää kuitenkin se, että vertailua tehdessä etsittävän merkkijonon aiheyhteys on tiedossa, mikä helpottaa etsintää huomattavasti, sillä kohdejoukkoa voidaan rajata. Jos tulkittava epäselvä käsiala on peräisin kauppalapulta, rajaa tämä vertailujoukon kaupassa myytäviin artikkeleihin ja todennäköisesti vielä tarkemmin lähes päivittäin ostettaviin tavaroihin. Jos taas etsittävä merkkijono on osittainen osoite, tiedämme rajata kohdejoukon kartalta löytyviin kadun nimiin ja luultavasti johonkin tiettyyn kaupunkiin tai sen osaan.

Henkilön itsensä kirjaamille ammattinimikkeille vastineita etsittäessä kohdejoukon rajaus edellä mainittujen esimerkkien tapaan ei onnistu. Kohdejoukkona toimivat kaikki viralliset ammattinimikkeiden kirjoitusmuodot mutta tämän enempää emme sitä saa rajattua. Lisäksi osassa tapauksista etsittävää merkkijonoa ei edes ole kohdejoukossa eikä tätäkään voida tietää etukäteen vaan nimikkeestä riippumatta vastinetta on etsittävä kohdejoukosta.

3.1 SQL-kielen tarjoamat työkalut ongelman käsittelyyn

Ohjelmointikielissä on yleensä valmiina yksinkertaisia, sisäänrakennettuja merkkioperaattoreita merkkijonojen vertailuun ja muokkaamiseen. SQL-kielessä merkkijonojen vertailuun käytössä ovat suora vertailu, jonka avulla kyselyssä voidaan palauttaa ne rivit, jotka täsmäävät merkki merkiltä annettuun hakuehtoon, ja *like*, jonka avulla voidaan vertailla osia merkkijonosta käyttämällä hakuehtona annettavan merkkijonon yhteydessä jokerimerkkiä. Näistä jokerimerkeistä yleisimmin käytetyt ovat prosenttimerkki (%), jota edeltävät tai seuraavat merkit voivat olla mitä tahansa ja niiden lukumäärä samoin mikä tahansa, sekä alaviiva (_), jonka tilalle hyväksytään mikä tahansa yksi merkki.

Suoralla vertailulla saadaan palautettua tässä työssä sellaiset henkilöiden ilmoittamat ammattinimikkeet, jotka täsmäävät täysin Tilastokeskuksen ammattinimikkeisiin. Katkaisemalla hakuehto prosenttimerkillä, ja täten jättämällä haettu merkkijono esimerkiksi muotoon ”%siivooja”, taas saadaan tuloksena kaikki siivooja-loppuiset ammattinimikkeet, olivatpa he sitten laitos-, teollisuus tai myymäläsiivoojia. Näin yhdistelemällä hakuehtojen jokerimerkkejä saadaan esimerkiksi hakuehdolla ”like ’%asteen _ehtori’” saadaan kaikki ylä- ja ala-asteiden lehtorit sekä rehtorit. Näistä suora vertailu on luonnollisesti ainoa tapaus, jolla henkilön ilmoittamalle ammattinimikkeelle saadaan suoraan vastine virallisista ammattinimikkeistä ja henkilölle saadaan sitä kautta lajiteltua ammattinimikkeen sisältämään ammattiluokkaan.

Muille tapauksille joudutaan tapauksesta riippuen luomaan erilaisia käsittelysääntöjä, funktioita ja proseduureja, jotta vastine ammattinimikkeelle löytyy. Yksinkertaisimmillaan tämä voi tarkoittaa ammattinimikkeen yhteydessä olevan lyhenteen avaamista, jonka jälkeen synonyymi löytyy suoraan. Esimerkiksi muokkaamalla henkilöiden ammattinimikkeitä siten, että kaikki ”lääk.”-loppuiset nimikkeet muunnetaan ”lääkäri”-

loppuisiksi saadaan ammattinimikkeelle ”eläinlääk.” heti vastine ”eläinlääkäri”. Toisen ääripäänä voisi olla tapaus, jossa ammattinimikettä on lyhennetty, se on kirjoitettu väärin sekä ammatin kuvauksessa on käytetty vierasperäistä tai puhekielessä vakiintunutta sanaa, joka eroaa kirjakielestä. Tällaisen tapauksen käsittelyssä tarvitaan jo useita käsittelysääntöjä ja apuvälineitä ennen kuin ammattinimike saadaan sellaiseen muotoon, että pari virallisista ammattinimikkeistä sille löydetään.

3.2 Valmiit ratkaisut

Valmiita SQL-kielellä toteutettuja kokonaisuuksia, jotka soveltuisivat merkkijonon lajitteluun valmiiksi annettuihin vastineisiin, en löytänyt ainoatakaan. Kaupallisia sovelluksia samankaltaisiin ongelmiin on kyllä olemassa, mutta niiden ongelmana on ”kielimuuri”, eli englanninkielen käsittelyyn tarkoitetut ohjelmat eivät yleensä anna kovin hyviä tuloksia käytettäessä suomenkielisiä merkkijonoja. Tällöin yksinkertaiset vertailut toki toimivat samalla tavalla kielestä riippumatta, mutta lisätyt käsittelyt eivät toimi halutulla tavalla. Esimerkiksi sanojen ääntämiseen ja taivuttamiseen muodostetut käsittelyt voivat tuottaa jopa täysin vääriä tuloksia sekä ymmärrettävästi suomen kielen sijamuodot ovat tällaiselle ohjelmalle täysin vieraita.

3.3 Oman ratkaisun rakentamisen etuja

Erittäin tärkeä syy oman järjestelmän tekemiseen on edellä sivuttujen lisäksi järjestelmän muokattavuus ja ylläpidettävyys. Itse laaditun järjestelmän rakentamisen hyötyjä verrattuna valmiiseen ratkaisuun ovat esimerkiksi:

- Uusien käsittelysääntöjen ja lyhenteiden lisääminen on helppoa.
- Reunaehtojen säätäminen (esimerkiksi ikärajat eläkeläisille) on toteutettavissa yhden parametrin muuttamisella.
- Virheellisten synonyymien korjaaminen on mahdollista käymällä läpi saatuja tuloksia ja päivittämällä virheelliset synonyymit oikeiksi.
- Uusien synonyymien lisääminen onnistuu yhdellä *insert*-lauseella, jolla synonyymitauluun voidaan lisätä joko yksi tai useampia synonyymejä kerrallaan.

- Nimikkeiden, joille ei ole löytynyt vastinetta, tarkastelu on helppoa ”roskakorin” ansiosta, jonne löytymättömät ammattinimikkeet päätyvät.
- Toiminnallisuuden lisääminen tulevaisuudessa havaittujen puutteiden perusteella onnistuu lisäämällä vaiheita ammattinimikkeiden käsittelyketjuun.

Työn toteuttaminen SQL-kielellä mahdollistaa perimmiltään saman toteutuksen hyödyntämisen myös mahdollisissa muissa merkkijonojen vertailuun liittyvissä ongelmissa. Perustoiminnallisuus voidaan helposti erottaa yksinomaan ammattinimikkeiden vertailuun toteutetuista käsittelysäännöistä, jolloin omia, tiettyyn kokonaisuuteen liittyviä käsittelysääntöjä luomalla järjestelmä voidaan monistaa käsittelemään erilaisia merkkijonoja. Järjestelmän hallinnoinnin tapahtuessa täysin yrityksen sisällä, on edellä mainittujen muutosten ja lisäysten teko hyvin yksinkertaista ja järjestelmän toiminnallisuus läpinäkyvää, sillä ammattinimikkeiden käsittelyssä käytettävä lähdekoodi funktioineen ja proseduureineen on kaikkien yrityksen työntekijöiden nähtävissä. Yrityksen kaikki poiminnat eri tietokannoista tehdään käyttämällä SQL-kieltä, joten jo sen takia järjestelmän toteuttaminen käyttäen SQL-kieltä oli hyvin luonnollista.

4 Toteutuksen suunnittelu

Suunnittelun apuna toimi Bisnode Finland Oy:n kehityspäällikkö Matti Jokela. Hän on ollut mukana 1990-luvun alussa kehittämässä tunnistusjärjestelmää, jonka avulla henkilöitä voidaan tunnistaa nimi- ja osoitetietoja käyttäen väestötietojärjestelmää vasten. Tässä hyvin monimutkaisessa järjestelmässä henkilölle etsitään annettujen tietojen perusteella vastinetta käyttäen avuksi muun muassa kirjoitusvirheiden korjausta, jatkuvasti kerättäviä katunimien synonyymejä ja henkilöiden entisiä nimiä ja osoitteita.

4.1 Viitteet tunnistusjärjestelmään

Yhteneväisyydet tunnistusjärjestelmään tavassa täsmätä merkkijonoja eri keinoin olivat ilmeiset, joten aluksi kävimme Jokelan kanssa palavereja siitä, miten työssä kannattaisi lähteä liikkeelle ja mitä tunnistusjärjestelmässä käytetyistä menetelmistä voisi mahdollisesti käyttää myös tässä työssä. Tunnistusjärjestelmässä vastineita etsitään ammattinimikkeistä poiketen usealle eri merkkijonolle kuten nimi, osoite ja postinumero. Täl-

löin kohdejoukkoa voidaan vuorotellen rajata eri elementtien avulla ja löytyneet ehdokkaat pisteytetään niiden paremmuusjärjestykseen saattamiseksi. Ammattinimikkeitä vastineita haettaessa vastaavan kohdejoukon rajaamisen ollessa vaikeaa pitäisi järjestelmästä saada toteutukseltaan niin kevyt, että kohdejoukkoa voitaisiin käyttää sellaisenaan.

Tunnistusjärjestelmässä käytössä on myöhemmin esiteltävä Levenshteinin algoritmi, jolla voidaan tutkia merkkijonojen samankaltaisuutta. Myös muita ideoita ja toiminnallisuutta omaan järjestelmäni toteuttamiseen tuli suoraan tunnistusjärjestelmästä. Vaikka toiminnallisuutta ei ollut suoraan siirrettävissä järjestelmien olennaisista eroista johtuen, hyväksi havaitut toteutuksen osat oli järkevää monistaa käyttöön omaa järjestelmää ajatellen.

4.2 Projektiryhmä

Projektia varten perustettiin viiden hengen projektiryhmä antamaan vuosikymmenten kokemuksen perusteella mielipiteitä ja näkemyksiä toteutuksesta sekä pitämään kirjaa projektin etenemisestä. Projektiryhmä kokoontui parin viikon välein, ja näissä kokouksissa käytiin läpi tehdyt toimenpiteet, aikataulutettiin tulevat toimenpiteet ja käsiteltiin mahdolliset toiminnallisuudesta heränneet kysymykset.

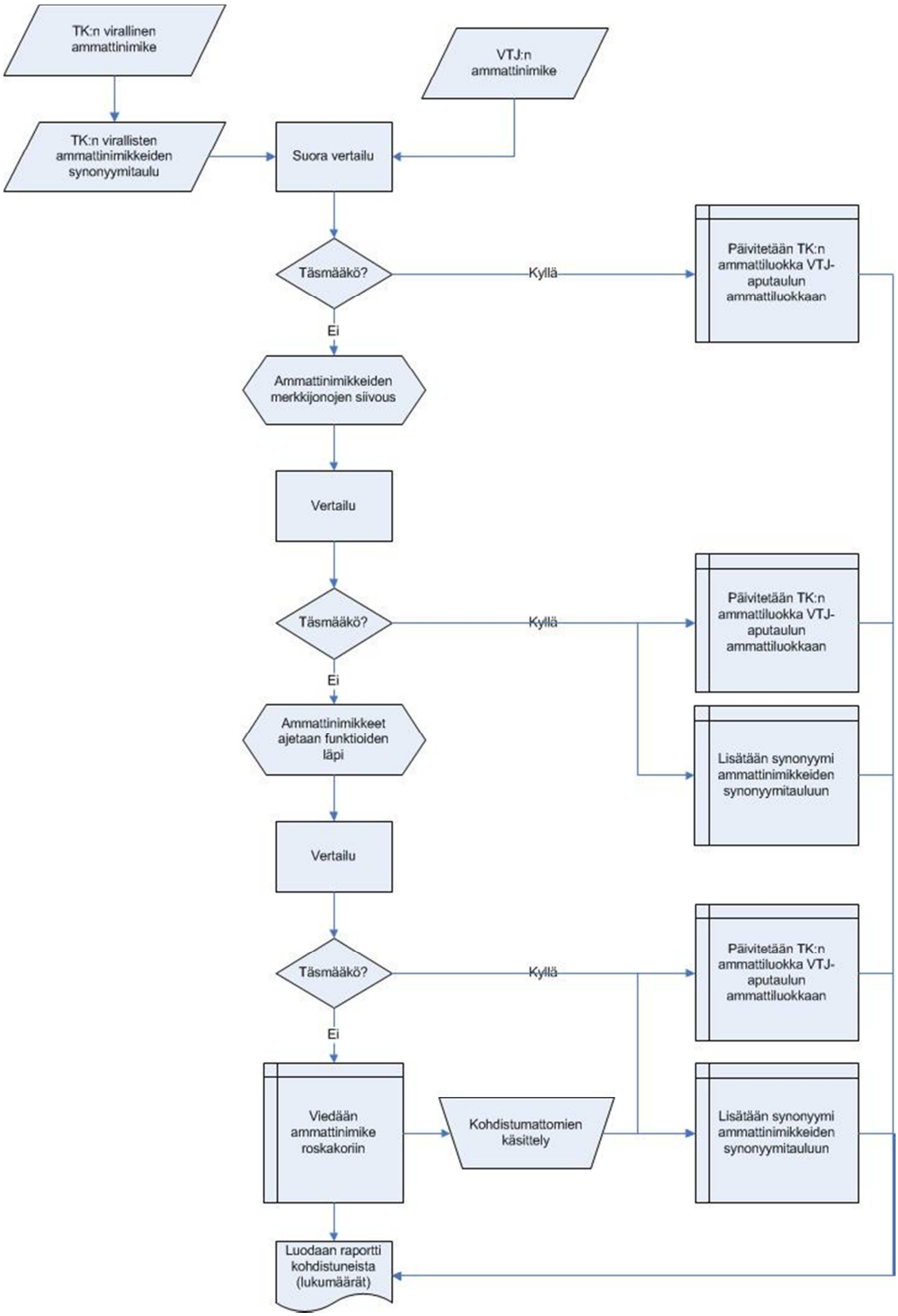
Aikaisessa vaiheessa sovittuja asioita projektiryhmän kesken oli muun muassa se, että alaikäisten henkilöiden ammattinimikkeille ei lähdetä hakemaan vastinetta virallisista nimikkeistä, sillä heitä ei kuitenkaan käytetä kohderyhmiä poimiessa, sillä tämä on katsottu kuluttajansuojalaissa hyvän tavan vastaiseksi (5).

Sovittiin myös eläkeläisen ikärajaksi 68 vuotta, joten kaikki yli 68-vuotiaat merkataan eläkeläisiksi riippumatta ammattinimike-kentän sisällöstä. Osasyy tähän päätökseen oli se, että vanhemmat ihmiset yleensä muuttavat harvemmin kuin nuoremmat ja koska ylivoimaisesti suurin osa ammattinimike-ilmoituksista tulee Väestörekisterikeskuksen tietopalveluasiantuntija Kristiina Heikkilän mukaan henkilön ilmoittaman muuttoilmoituksen kautta. Tämän ikäisellä henkilöllä saattaa olla väestötietojärjestelmässä kirjattuna useamman vuoden tai jopa vuosikymmenien takainen ammattinimike, jonka oikeellisuudesta ei ole enää mitään varmuutta.

Alustavasti, puhtaasti käsin tekemässäni eri ammattinimikkeiden läpikäynnissä huomasin, että hyvin moni on syöttänyt ammattinimike-kenttään useamman ammatin. Näiden tapausten käsittelystä sovittiin, että järjestelmä yrittää löytää vastineen mille tahansa henkilön syöttämistä ammattinimikkeistä, sillä mitään varmuutta siitä, että esimerkiksi ensimmäinen henkilön syöttämistä ammattinimikkeistä olisi niin sanotusti ensisijainen ammatti, ei ole.

4.3 Järjestelmän toiminta

Näiden asioiden selkiydyttyä laadin suunnittelun avuksi ensin kuvauksen tavasta, jolla ammattinimikkeitä käsiteltäisiin prosessin eri vaiheissa, joka esitetään kuvassa 1. Kuvan tarkoituksena oli havainnollistaa sekä projektiryhmälle että itselleni, miten järjestelmä tulisi toimimaan. Samalla sain uusia ideoita esimerkiksi siitä, miten löydetty ammattinimikesynonyymi liitettäisiin haetulle nimikkeelle.



Kuvio 1. Järjestelmän toiminta

4.3.1 Perustoiminnallisuus

Kuvaus helpotti hahmottamaan, miten toiminnallisuus tulisi suunnilleen toteuttaa. Järjestelmä rakennettaan täysin erilliseksi, jolloin sen ollessa valmis järjestelmä voidaan liittää suoritettavien ylläpitoajojen listaan muiden jatkoksi. Näin ollen ammattinimikkeiden käsittelyjärjestelmä ei ole riippuvainen muista ylläpitoajoista eikä myöskään muut ylläpitoajot tästä.

Samaten ammattinimikkeiden varsinainen muokkaus ja synonyymien hakeminen järjestelmän sisällä toteutetaan siten, että kaikki ammatit viedään omaan käsittelytauluunsa, johon varataan lisäksi paikka viralliselle ammattinimikkeelle sekä sen koodille. Tähän tauluun järjestelmä luo lisää sarakkeita tarvittaessa esimerkiksi auki kirjoitetuille lyhenteille ja muokatuille nimikkeille, joten lopuksi ammattinimikkeen uusi koodi voidaan yhdistää tästä väliaikaisesta käsittelytaulusta varsinaiseen poimintatauluun. Näin järjestelmä toimii täysin omana prosessinaan koko sen suorituksen ajan. Vasta poimintataulua päivitettäessä tarvitaan ulkopuolista, kaikkien käytössä olevaa taulua.

Toinen prosessissa päivitettävä taulu on varsinainen synonyymitaulu, johon kerätään jatkuvasti uusia synonyymejä ammattinimikkeiden virallisille nimikkeille. Tämän taulun ollessa kattava löytyy useammalle ammattinimike suoraan tästä taulusta. Tällöin ammattinimikettä ei tietenkään tarvitse käsitellä raskaampien käsittelyjen kautta tehden järjestelmän ajon jatkossa aina kevyemmäksi. Lisäksi mahdollisuutena on käyttää jo luotuja synonyymejä synnyttämään edelleen uusia synonyymejä, vaikka tätä ei aiotakaan aluksi käyttää, jotta vältyttäisiin turhilta vääriltä synonyymeiltä. Esimerkkinä tällaisesta tapauksesta henkilö, jonka ammatti on postinkuljettaja, olisi saattanut kirjoittaa ammattinsa virheellisesti muodossa ”possinkuljettaja”. Näiden kirjoitusmuotojen editointietäisyys on yksi, joten synonyymipari lisättäisiin synonyymitauluun. Kuitenkin sallittaessa edelleen editointietäisyyden avulla lisätyn synonyymin käyttö uusien synonyymien muodostuksessa, olisi merkkijonojen ”possinkuljettaja” ja ”lossinkuljettaja” editointietäisyys niin ikään yksi, jolloin virheellisesti merkattaisiin ”lossinkuljettaja” synonyymiksi postinkuljettajalle.

Kaksi muuta järjestelmän tuottamaa taulua tallennetaan myös erilleen kaikesta muusta toiminnallisuudesta. Näistä raporttitaulu luodaan uudelleen joka ajon yhteydessä ja sen sisältö on hyvin yksinkertainen: taulu sisältää sarakkeet selitettä ja lukumäärää varten

ja taulua päivitetään ajon edetessä, jolloin talteen saadaan lukumäärät sekä ennen että jälkeen eri vaiheita. Toinen lisätauluista on eräänlainen roskakori, jonne viedään ajon jälkeen sellaiset ammattinimikkeet, joille ei löydetty vastinetta virallisista nimikkeistä ja joita ei roskakorista vielä löydy. Nämä jäävät toistaiseksi käsin läpikäytäviksi ja mahdollisten uusien käsittelysääntöjen tai toiminnallisuuksien lisäämisen jälkeen roskakorissa olevia ammattinimikkeitä voidaan yrittää kohdentaa uudelleen viralliseen ammattinimikelistaukseen.

4.3.2 Järjestelmän suorituksen eri vaiheet

Eri vaiheissa sillä hetkellä ilman synonyymiä olevat ammattinimikkeet viedään aina omaan aputauluunsa käsittelyä varten ja vaiheen suorituksen jälkeen niiden ammattinimikkeiden, joille synonyymi löydettiin, viralliset ammattinimikkeet päivitetään aluksi luotuun käsittelytauluun. Vaiheet synonyymien löytämiseksi etenevät progressiivisesti niin, että ensimmäiset vaiheet ovat käsittelyn osalta kevyitä toteuttaa ja jälkimmäiset käsittelyt ovat huomattavasti raskaampia suorittaa. Näin saadaan aluksi hyvin yksinkertaisilla vertailuilla ja toimenpiteillä pienennettyä kohdejoukkoa merkittävästi ja myöhempiä, raskaimpia käsittelyjä ei tarvitse tehdä niin isolle joukolle ammattinimikkeitä.

4.3.3 Toimintaympäristö

Työn toteuttamisessa käytetään puhtaasti SQL-kieltä ja sen avulla toteutettuja funktioita ja proseduureja. Nämä ovat yrityksessä normaaleissa tuotantotöissä käytettyjä välineitä, joten ne ovat itselle varsin tuttuja ja valmiiksi saatavilla. SQL-kieli ei ole varsinainen ohjelmointikieli vaan ennemminkin kyselykieli, jonka ominaisuuksia tässä työssä käytetään hyväksi: yksinkertaistettuna voitaisiin ajatella, että jokaiselle erilaiselle ammattinimikkeelle pyritään kyselyn tuloksena palauttamaan tasan yksi tulos virallisista ammattinimikkeiden kirjoitusmuodoista ja tästä ammattinimikeparista muodostuu uusi synonyymipari.

SQL-kielellä on toki mahdollista tehdä hyvin toimivaa ohjelmointia funktioiden ja proseduurien muodossa, jotka ovat tässäkin järjestelmässä suuressa osassa. Mikäli järjestelmään on tarpeen lisätä toiminnallisuutta ohjelmointikielistä, voidaan niitä tuoda SSIS-paketin sisään rakennettuina esimerkiksi C# -ohjelmointikieltä käyttäen, joka on

täysimittainen olio-ohjelmointikieli. SSIS -paketti on erityisen kehittimen avulla luotu kokonaisuus, joka suorittaa sen sisään määritellyt tehtävät SQL-ympäristössä ennalta määritellyssä järjestyksessä ja voi sisältää hyvin erilaisia vaiheita kuten esimerkiksi tiedostojen lukua tietokantaan, kovalevyllä sijaitsevien tiedostojen käsittelyä tai FTP-siirtoja. Toistaiseksi tämän työn toteuttamiseksi ei tarvittu piirteitä varsinaisista ohjelmointikielistä, mutta on hyvä tietää, että mahdollisuus tähän on tarvittaessa olemassa. Järjestelmä rakennetaan ympäristössä, jonka tietokannan hallintajärjestelmänä on Microsoft SQL Server ja kehittimenä käytössä on Microsoft SQL Server Management Studio.

4.4 Integrointi ylläpitoajoihin

Toimiessaan järjestelmä käsittelisi ammattinimikkeet samalla syklillä kuin muutkin tuotantoaineistoa työntekijöiden käyttöön valmistelevat prosessit eli niin sanotut ylläpitoajot. Nämä ylläpitoajot ovat yleensä viikonloppuisin suoritettavia jobeja, jotka käynnistetään sen jälkeen, kun Väestörekisterikeskukselta on saatu uusi väestötietojärjestelmän kanta, joka sisältää aina tuoreimmat tiedot eli käytännössä kannan päivitettynä kuluneen viikon aikana tapahtuneilla muutoksilla.

Ylläpitoajojen tuloksena syntyy lukuisia tauluja, joista löytyvät poimintoja varten tarvittavat tiedot koskien esimerkiksi henkilöitä, rakennuksia, rakennuslupia, muuttohistorioita ja kunta-, maakunta- ja läänitietoja. Näistä koostetaan lopuksi erillinen poimintataulu, johon on liitetty eri tauluista yleisimpiä poiminnoissa käytettäviä poimintaehtoja. Tähän poimintatauluun tultaisiin liittämään uutena sarakkeena uusi ammattinimikkeen saama Tilastokeskuksen koodi.

5 Toiminnallisuuden toteuttaminen

5.1 Lähtökohdat

Järjestelmässä lähdetään liikkeelle erikseen poimintoja varten luodusta aputaulusta, johon on otettu mukaan kaikki Suomen väestön markkinointikelpoiset henkilöt. Markkinointikelpoisella väestöllä tarkoitetaan sitä osaa ihmisistä, joita voidaan lähestyä osoitteellisen suoramarkkinoinnin keinoin, joten pois on jätetty kaikki ulkomailla asuvat ja osoitteettomat henkilöt sekä henkilöt, jotka ovat asettaneet itselleen joko suoramarkkinointi- tai turvakiellon. Markkinointikelpoisia henkilöitä maaliskuussa 2011 oli noin 5 070 000 kappaletta.

5.2 Yleisrakenne

Aloitin toteutuksen luomalla kaksi taulua. Ensimmäiseen, henkilötauluun, vein kaikkien VTJ:stä löytyvien henkilöiden henkilötunnuksen, ammattinimikkeen ja paikan viralliselle ammattinimikkeelle. Toiseen tauluun, synonyymitauluun, vein kaikki erilaiset ammattinimikkeiden kirjoitusmuodot ja niin ikään paikan viralliselle ammattinimikkeelle, josta löytyneitä ammattinimikkeitä voitaisiin viedä kiinni henkilötaulun henkilötunnuksiin.

Synonyymitaulun kasvattaminen aluksi oli tärkeää eri tavoin, jotta tauluun saataisiin oikeasti eri ammattinimikkeiden synonyymejä. Juuri luodussa synonyymitaulussahan ei nimittäin tietysti ole muuta kuin Tilastokeskuksen ammattinimikkeet, joiden synonyymipareina ovat samat merkkijonot. Sellaisten ammattinimikkeiden ilmentymien lukumäärien laskeminen VTJ:stä, joita ei suoralla vertailulla saatu osumaan Tilastokeskuksen ammattinimikkeisiin, antoi hyvä kuvan sellaisista nimikkeistä, jotka kannatti heti lisätä synonyymitauluun käsin.

Tuloksista tutkittiin ensisijaisesti sellaisia ammattinimikkeitä, joiden esiintymislukumäärä VTJ:ssä oli suuri. Näin saataisiin jäljelle jäävää löytämättömien joukkoa pienenemään nopeammin ja siten karsittua suurta massaa. Tällaisia, käsin lisättyjä synonyymejä, olivat esimerkiksi monet suuresti lyhennetyt, jopa vakiintuneet käsitteet kuten ”mv”, maanviljelijä, ”pph”, perhepäivähoitaja tai ”hml”, hammaslääkäri. Tällaisia ammat-

tinimikkeitä on pitkälti mahdotonta kohdistaa niiden oikeisiin kirjoitusmuotoihin koneellisella käsittelyllä.

Samalla lisättiin useita muita synonyymejä, joiden lisääminen heti alkuun tuntui tarpeelliselta. Nämä sisälsivät esimerkiksi englanniksi kirjoitettuja ammattinimikkeitä, sillä kuten aiemmin on mainittu, englanninkielistä ammattinimikelistausta ei ollut saatavilla, joten ne olivat lisättävä suomenkielisten ammattinimikkeiden synonyymeiksi.

Opiskelijoita, eläkeläisiä ja ammatittomia varten synonyymitauluun perustettiin omat nimikkeensä koodeineen, sillä nämä ryhmät halutaan erottaa muista. Luomalla oma koodi ja nimike ko. ryhmille ne voitaisiin helposti poimia erikseen väestötietojärjestelmästä.

5.3 Alaikäisten, eläkeläisten ja ammatittomien alustava karsiminen

Vertailtavan joukon pienentämiseksi päätin aluksi tehdä proseduurin, joka yksinkertaisella vertailulla merkkaa väestön alaikäiset, eläkeläiset sekä henkilöt, jotka eivät ole antaneet itselleen mitään ammattinimikettä. Projektiryhmän päätöksien mukaisesti alaikäisten ammattinimikkeitä ei lähdetä käsittelemään ja kaikki yli 68-vuotiaat henkilöt merkataan siirtyneiksi eläkkeelle ammatista riippumatta.

Näistä alaikäisten ja eläkeläisten merkitseminen tapahtui pelkästään laskemalla henkilön ikä henkilötunnuksesta, jonka avulla alaikäiset ja eläkeläiset merkattiin niitä varten varattuihin sarakkeisiin. Eläkeläisiin lajiteltiin tietysti suoraan myös henkilöt, jotka saatiin suoraan vertaamalla ammattinimikettä synonyymitaulun merkkijonoon ”eläkeläinen”. Ammatittomiksi merkattiin niin ikään yksinkertaisesti tässä vaiheessa ne henkilöt, joiden ammattinimikekenttä oli tyhjä.

Näiden toimenpiteiden jälkeen saatiin myös pohjaa tilastolle siitä, kuinka paljon erilaisia ammattinimikkeitä väestötietojärjestelmässä oli ja kuinka monelle henkilölle ammattinimike oli yritettävä löytää. Maaliskuussa 2011 markkinointikelpoisista henkilöistä alaikäisiä oli noin 1 050 000, eläkeläisiä noin 725 000 ja käsittelyn näkökulmasta ammatittomia myös noin 725 000 kappaletta.

5.4 Karsiminen suoralla vertailulla

Seuraava looginen ja yksinkertainen askel oli vertailla henkilön antamaa ammattinimikettä virallisiin kirjoitusmuotoihin suoralla merkkijonovertailulla. Tässä vaiheessa ei otettu kantaa mahdollisiin kirjoitusvirheisiin tai muuhun epämääräisyyteen henkilöiden ilmoittamissa ammattinimikkeissä, vaan yhdistettiin täysin täsmäävät nimikkeet niiden virallisiin versioihin ja saatiin näin vertailujoukkoa taas pienennettyä.

5.5 Karsiminen merkkijonon siivouksen avulla

Seuraavassa vaiheessa käytin hyväkseni tunnistusjärjestelmässä käytettyä funktiota, josta käytetään nimeä ”charclean”. Tämän funktion tehtävät ovat siivota merkkijonosta pois kaikki ei-kirjainmerkit sekä poistaa kaikki tuplakonsonantit ja -vokaalit, joiden oli tunnistusjärjestelmää kehitettäessä todettu aiheuttavan paljon kirjoitusvirheitä suomenkielessä. Funktion tuloksena esimerkiksi merkkijono ”1.sairaan-hoitaja” muuttuu muotoon ”sairanhoitaja”. Kun kaikki viralliset ammattinimikkeet oli muokattu saman funktion avulla, saatiin kyseessä oleva virheellisesti kirjoitettu kirjoitusmuoto täsmäämään suoraan viralliseen ammattinimikkeeseen. Vertailu merkkijonon siivouksen jälkeen tehtiin edelleen suorana vertailuna ja pienien kirjoitusvirheiden selvittäminen jätettiin tuleviin vaiheisiin. Vertailun jälkeen löytymättömiä henkilöitä jäi noin 110 000 kappaletta, joka tarkoittaa noin 35 000 erilaista ammattinimikkeen kirjoitusmuotoa.

5.6 Jatkotoimenpiteiden mietintä

Tutkimalla silmämääräisesti jäljelle jääneitä löytymättömiä ammattinimikkeitä havaitsin, että seuraavaksi tarvetta oli erinäisille proseduureille:

- Proseduuri, joka muokkaisi erilaisia kirjoitusmuotoja siitä, että henkilö on eläkkeellä, muotoon ”eläkeläinen”. Näin saataisiin ohitettua sellaiset tapaukset, joissa ammattinimikkeeksi olisi annettu esimerkiksi ”poliisi eläk.” eikä nimike yhdistyisi virheellisesti nimikkeeseen ”poliisi”. Samanlainen operaatio tehtäisiin samassa opiskelijoille, jotta tätä ryhmää saataisiin samoin karsittua ja päästäisiin paremmin käsiksi itse ammattinimikkeisiin, jotka eivät vielä osuneet virallisiin nimikkeisiin.

- Proseduuri, joka tyhjentäisi ammattinimikkeen kokonaan tietyiltä tapauksilta. Tällaisia ovat esimerkiksi ”entinen siivooja”, ”työtön” ja ”lapsi”, jotka eivät kerro mitään henkilön tämänhetkisestä ammatista ja ovat näin enemmän haitaksi kuin hyödyksi ammattinimikettä poimintaehtona käytettäessä.
- Proseduuri, joka poistaisi ammattinimikkeiden yhteydessä esiintyvät sellaiset määreet, jotka toimivat pelkästään tarkentavina varsinaiseen ammattinimikkeeseen eivätkä anna lisäarvoa itse nimikkeelle, ainoastaan vaikeuttavat virallisen ammattinimikevastineen koneellista etsintää. Tällaisia ovat esimerkiksi ”avustava”, ”vanhempi” tai ”päätoiminen”.
- Proseduuri, joka avaisi erilaisia lyhenteitä ja kirjainyhdistelmiä oikeaan kirjoitusasuunsa, jotta esimerkiksi ammattinimike ”eläinlääk.” saataisiin muotoon ”eläinlääkäri” tai ammattinimike ”sosiaalitt.” muotoon ”sosiaalityöntekijä”.
- Proseduuri, joka avaisi erilaiset lyhenteet koulutukseen liittyvistä titteleistä. Varsinaisesti ammattiin liittymättömät koulutusnimikkeet haluttiin myös syrjään oikeiden ammattinimikkeiden käsittelyn tieltä.

Seuraavaksi perehdymme tarkemmin näiden edellä pohdittujen proseduurien merkitykseen ja niiden toteutuksiin.

5.6.1 Proseduuri eläkeläisten karsimista varten

Eläkeläisten karsimista varten tehty proseduuri on periaatteessa suuri kokoelma käsittelysääntöjä, joiden avulla eri lailla kirjatut eläkeläiset saataisiin luokiteltua. Myös joissain ammattinimikkeissä esiintyy sana eläke (esimerkiksi ”eläkeneuvoja”), minkä takia käsittelysääntöjä luodessa piti huomioida, ettei niitä luokiteltaisi eläkeläisiksi. Muuten eläkeläisten karsiminen on helppoa, sillä eläkkeellä olemisen termi on niin sanotusti poissulkeva, jolloin sen maininta ammattinimikkeessä johtaa aina henkilön luokittelemiseen eläkeläiseksi. Suomenkielen sanojen taivutusmuodot tuovat oman haasteensa käsittelysääntöjen laatimiseen, sillä esimerkiksi eläkeläiseksi halutaan merkittävän niin ”eläkeläinen”, ”eläkkeellä” kuin myös ”eläköitynyt”. Käsittelysääntöjen laatiminen tehtiin pääosin käyttämällä SQL:n yleisiä vertailuehtoja ja yhdistelemällä niitä muodostamaan toimivia kokonaisuuksia.

5.6.2 Proseduuri epäkelpojen ammattinimikkeiden tyhjentämistä varten

Epäkelpojen ammattinimikkeiden tyhjentämiseen tehdyn proseduuriin käytettävät käsittelyehdot ovat hyvin hajanaiset. Palauttamalla haulla löytymättömiä ammattinimikkeitä eri järjestyksessä, kuten pituus- ja aakkosellisessa järjestyksessä, löytyi erilaisia tapauksia epäkelvoista nimikkeistä, jotka käsittelyehtoja sovellettaessa saataisiin keskitetysti tyhjennettyä. Esimerkiksi täysin sattumanvaraisilta näyttävien merkkijonojen käsittelyyn tehtiin oma käsittelysääntönsä, samoin kuin lähes kokonaan numeroita sisältävien merkkijonojen käsittelyyn. Yllättävän paljon löytyi myös nimikkeitä kuten ”vauva”, ”lapsi” ja ”rouva” -tyylisiä nimikkeitä, jotka tietysti ovat turhia. Omat käsittelysääntönsä saivat myös jo edellä mainitut nimikkeet, joihin liittyy esimerkiksi määre ”entinen”, joka vie pohjan sen yhteydessä annetulta nimikkeeltä. Suuri osa epäkelpojen ammattinimikkeiden tyhjentämiseen laadituista käsittelysäännöistä on kuitenkin täysin ammattinimikkeitä silmäilemällä havaittuja sääntöjä.

5.6.3 Proseduuri ammattinimikettä tarkentavien osien tyhjentämistä varten

Käsittelysäännöt ammattinimikettä tarkentavien osien tyhjentämiseen ammattinimikekentästä ovat niin ikään pitkälti käsin läpikäymällä havaittuja. Vaikka näitä käsittelysääntöjä ei ole läheskään niin paljon kuin muissa proseduureissa, on niistä suuri apu. Jatkokäsittelyssä ymmärrettävästi ei voida hyväksyä kovinkaan monen merkin eroavaisuutta henkilön syöttämän ammattinimikkeen ja virallisen nimikkeen välillä, joten tämän lainen nimikettä tarkentava määre voi aiheuttaa sen, ettei ammattinimikkeelle löydetä vastinetta lainkaan, vaikka nimike olisi muuten oikein kirjoitettu.

5.6.4 Proseduuri lyhenteiden avaamista varten

Lyhenteitä avaavan proseduurin käsittelysääntöjä kertyi kaikista eniten. Erilaisten sanojen, oli se sitten ammattia tarkentava alkuosa (esimerkiksi ”röntgen”) tai loppuosa, varsinainen ammatti (esimerkiksi ”lääkäri”), lyhentäminen on todella yleistä. Edellä esimerkkinä annettu ”röntgenlääkäri” olisi esimerkiksi kovin vaikea kohdistaa sen viralliseen kirjoitusmuotoon, mikäli se olisi lyhennetty äkkiseltään hyvinkin epämääräiseltä vaikuttavaan muotoon ”rntgnlääk”.

Yritin etsiä tähän käyttöön valmista listaa erilaisista sanojen lyhenteistä, sillä suuri osa ammattinimikkeissä esiintyvistä sanoista on kuitenkin yleistä suomenkielen sanastoa. Tarkoitukseeni sopivaa listaa ei kuitenkaan joutunut, joten päädyin tekemään sellaisen itse.

Hyvä keino yleisimpien lyhenteiden löytämiseksi oli tässäkin tapauksessa suorittaa kyselyitä, joissa eri ammattinimikkeiden, joille ei vielä ole löydetty synonyymiä, kirjoitusmuotojen ilmentymiskerrat väestötietojärjestelmässä lasketaan ja tuloksena palautetaan eri kirjoitusmuodot järjestettynä laskevassa järjestyksessä ilmentymiskertojen suhteen. Tuloksia läpi käymällä erottuvat yleisimmät kirjoitusmuodot ja niiden mukana yleisimmät lyhenteet. Selvää on, että myös lyhenteiden avaaminen eri nimikkeistä helpottaa suuresti ammattinimikeparien etsintää suurimpien merkkijonojen välisten eroavaisuuksien poistuessa lyhenteen avauksen yhteydessä.

5.6.5 Proseduuri koulutustitteleitä varten

Aluksi oli epäselvää, mihin luokkaan erilaiset koulutukseen liittyvät tittelit lajiteltaisiin, sillä niillä ei varsinaisesti ole mitään tekemistä henkilön ammatin kanssa. Projektiryhmän kokouksessa päätettiin, että näille luotaisiin oma luokkansa ja siihen omat nimikkeensä yleisimmille titteleille. Tämä katsottiin hyödyllisemmäksi kuin kyseisten titteleiden omaavien henkilöiden ammattinimikkeen tyhjentäminen kokonaan, sillä oma luokka mahdollistaisi kohderyhmän rajaamisen myös tarvittaessa käyttäen apuna koulutusnimikkeitä. Myös tässä tapauksessa etsin valmista listaa yleisistä koulutusnimikkeistä ja niiden lyhenteistä. Pohja koulutusnimikkeiden synonyymitaululle, jossa luetellaan yliopistotutkinnot lyhenteineen, löytyi Wikipediasta (6). Tätä listaa laajentamalla sain hyvän synonyymitaulun koulutustitteleiden osalta lisäämällä siihen muita koulutuksiin liittyviä titteleitä ja tutkintoja, sekä lisäämällä käsin muutamia yleisiä synonyymejä, kuten ”ins.” insinöörille ja ”di” diplomi-insinöörille. Tämän jälkeen koulutusnimikkeitä varten kehitettiin niin ikään oma proseduurinsa käsittelysääntöineen.

Erilaisia käsittelysääntöjä, joiden mukaan ammattinimikkeiden kirjoitusmuotoja muokataan ja käsitellään, kertyi tähän asti muodostetuille proseduureille yhteensä 772 kappaletta.

5.6.6 Ammattinimikkeiden vertailu proseduurien suorittamisen jälkeen

Edellä mainittujen proseduurien suorittamisen jälkeen muokatut ammattinimikkeet ajettiin taas ”charclean”-funktion läpi ja verrattiin virallisiin ammattinimikkeisiin. Erikseen jokaisen proseduurin suorittamisen jälkeen vertailua ei tehty, sillä mukana on ammattinimikkeiden kirjoitusmuotoja, jotka hyötyvät useamman edellä mainitun proseduurin suorituksesta (esimerkiksi ”väliaikainen teollisuussiiv.”, josta sekä poistetaan määre ”väliaikainen” että avataan lyhenne ”siiv.” siivoojaksi), joten näin hyöty kaikista tässä vaiheessa suoritettavista proseduureista saatiin yhdessä vertailussa kaikkien niiden suorituksen jälkeen. Kaikki tässä vaiheessa löytyneet vastineparit lisättiin alussa tehtyyn synonyymitauluun, jolloin ajettaessa vertailua uudelleen ei samoja henkilöiden syöttämiä ammattinimikkeitä tarvitse ajaa uudelleen muokkauksen läpi vaan vastineet löytyvät ensimmäisessä vaiheessa, suorassa vertailussa synonyymitauluun. Huomionarvoista näissä proseduureissa on se, että jokaiseen voidaan todella helposti lisätä käsittelysääntöjä jälkeenpäin. Tämä on hyvin tarpeellista varsinkin ensimmäisten ajokertojen läpi kohdennettavan ammattinimikejoukon pienentyessä, jolloin tarpeet uusille käsittelysäännöille nousevat selvemmin esiin.

5.7 Levenstein-algoritmi

Seuraavaksi käytin hyväkseni algoritmia, joka laskee annetuille kahdelle merkkijonolle niin sanotun Levenšteinin etäisyyden eli editointietäisyyden. Tämän kahden merkkijonon välisen editointietäisyyden laskevan algoritmin esitti venäläinen matemaatikko Vladimir Levenštein vuonna 1965. Algoritmi kuvaa sitä, montako operaatiota tarvitaan, jotta annettu merkkijono saadaan muutettua toiseksi. Editointietäisyyttä laskettaessa sallittuja operaatioita ovat merkin lisääminen, poistaminen ja korvaaminen toisella (7). Editointietäisyys sekä useat muut samankaltaiset algoritmit ovat nykyään laajalti käytössä. Hyvä esimerkki tästä on esimerkiksi lääketieteen saralla DNA:n proteiiniketjujen vertailu keskenään (8).

Editointietäisyyden laskemiseen tarvittava lähdekoodi on avoin ja täten jokaisen saatavilla esimerkiksi Internetistä. Toimintaperiaatteen ollessa julkinen on algoritmin toteutuksesta olemassa suuri määrä erilaisia toteutuksia, joiden tehokkuus vaihtelee suuresti. Yrityksemme käytössä oli ennestään yksi toteutus Levenstein-algoritmin toteutukses-

ta, joka on kirjoitettu SQL:n funktioksi ja sitä kutsutaan antamalla sille parametreina kaksi merkkijonoa, joiden välisen editointietäisyyden funktio palauttaa. Alla on esitetty yksinkertainen kysely, joka palauttaa kahden annetun merkkijonon, ”ympäristönsuunnittelija” ja ”ympäristösuunnittelija”, editointietäisyyden, joka on yksi.

```
select
sievaant.dbo.udf_levenshtein('ympäristönsuunnittelija','ympäristösuunnittelija')
```

SQL:n funktiot ottavat parametreina vastaan myös taulun sarakkeita, jotka mahdollistavat joko yhden merkkijonon vertailun jokaiseen taulun tietyssä sarakkeessa olevaan merkkijonoon tai suoraan kahden eri sarakkeen vertailun keskenään. Seuraavassa kyselyssä verrataan merkkijonoa ”ympäristönsuunnittelija” esimerkkitaulusta löytyviin kymmeneen samantapaiseen merkkijonoon.

```
select 'ympäristönsuunnittelija' as ympäristönsuunnittelija, ammatti as
synonyymiehdokas,
sievaant.dbo.udf_levenshtein('ympäristönsuunnittelija',ammatti) as levenshtein
from esimerkki_taulu
```

Edellä esitetyn kyselyn tulokset on esitetty kuvassa 2.

ympäristönsuunnittelija	synonyymiehdokas	levenshtein
ympäristönsuunnittelija	ympäristönsuojelun tarkastaja	13
ympäristönsuunnittelija	ympäristö- ja turvallisuusjoht	18
ympäristönsuunnittelija	ympäristön suoj tarkastaja	12
ympäristönsuunnittelija	ympäristönsuojeluasiamies	12
ympäristönsuunnittelija	ympäristönsuunnittelija	0
ympäristönsuunnittelija	ympäristönsuojelun professori	14
ympäristönsuunnittelija	ympäristönsuojelusiht ii rak t	14
ympäristönsuunnittelija	ympäristösuojelupäällikkö	13
ympäristönsuunnittelija	ympäristönsuojelupäällikkö	12
ympäristönsuunnittelija	ympäristönsuojelu sihteeri	11

Kuvio 2. Esimerkkikyselyn palauttamat tulokset

Editointietäisyys voidaan laskea myös taulukointina. Taulukossa 1 on esitetty kahden sanan, ”handball” ja ”ballad” editointietäisyys. Taulukon jokainen alkio vastaa sitä operaatioiden lukumäärää, joka tarvitaan siihen asti vaaka- ja pystyriveiltä löytyvien merkkijonojen muuttamiseksi samanlaisiksi. Merkkijonojen lopullinen editointietäisyys saadaan taulukon lävistäjästä eli jos merkataan A = ballad, B = handball on editointietäi-

syys D(A,B) = d_{mn} = $d_{6,8}$ = 6. Jos taas sanat olisivatkin ”hand” ja ”ball”, voidaan niiden editointietäisyys nähdä taulukon alkiosta, joka on kuudennen rivin kuudennesta sarakkeesta. Tällöin editointietäisyydeksi saadaan kolme, sillä vain a-kirjain on merkkijonojen kesken sama ja samalla paikalla, muut kirjaimista joudutaan vaihtamaan, jotta merkkijonot olisivat samat.

Taulukko 1. Editointietäisyyden laskeminen taulukoinnin avulla

d		h	a	n	d	b	a	l	l
	0	1	2	3	4	5	6	7	8
b	1	1	2	3	4	4	5	6	7
a	2	2	1	2	3	4	4	5	6
l	3	3	2	2	3	4	5	4	5
l	4	4	3	3	3	4	5	5	4
a	5	5	4	4	4	4	4	5	5
d	6	6	5	5	4	5	5	5	6

Yksittäisen editointietäisyyden palauttaminen on nykyisille tehokkaille palvelimille toteutustavasta riippumatta kevyt operaatio, jonka tulos palautetaan sekunnin murto-osassa. Sarakkeiden välisten editointietäisyyksien laskeminen on luonnollisesti raskaampaa sitä mukaa kuin sarakkeiden rivimäärät lisääntyvät, mutta edelleen kyse on useamman yksittäisen editointietäisyyden laskemisesta peräkkäin.

Ammattinimikkeiden kirjoitusmuotoja toisiinsa vertaillessa törmätään kuitenkin ongelmaan, sillä yksittäisen kirjoitusmuodon vastineen etsimisessä suuresta vertailujoukosta ei vertailujoukkoa voida rajata. Tällöin joudutaan vertailemaan yhtä merkkijonoa synonyymitaulun sarakkeeseen, joka sisältää kaikki erilaiset kirjoitusmuodot eri ammat-

tinimikkeistä. Jos etsittäviä ammattinimikkeiden kirjoitusmuotoja on esimerkiksi tuhat ja synonyymejä 100 000 saadaan vertailun tuloksena 1 000*100 000 = 100 000 000 arvoa. Näin suuren tulosjoukon laskeminen ei onnistu ilman, että suoritukseen varattaisiin useita päiviä eikä kaikkien mahdollisten parien editointietäisyyksien laskeminen näin ollen ole järkevää tai käytännöllistä.

5.7.1 Levenstein-algoritmien tehokkuuden vertailu

Etsin Internetistä useampia SQL-kielelle tehtyjä toteutuksia Levenstein-algoritmista, joita ryhdyin vertailemaan keskenään. Testauksessa käytin aluksi sadan valmiin synonyymin otosta johon syöttämääni merkkijonoa verrattiin ja jokaiselle parille palautettiin niiden editointietäisyys. Tämän suuruisilla vertailujoukoilla suuria eroja ei syntynyt, mutta oli ilmeistä, että editointietäisyyden laskemisessa käytettyjen lähdekoodien tehokkuuksissa oli eroja. Laajentamalla ensin synonyymien määrää tuhanteen ja lopulta kymmeneen tuhanteen alkoi eroja syntyä. Lopullisen testauksen suoritin kursorikäsittelyllä, jossa sadalle satunnaisesti poimitulle ammattinimikkeen kirjoitusmuodolle laskettiin editointietäisyys jokaisen synonyymitaulusta löytyvän synonyymin kanssa. Kursorikäsittely on eräänlainen toistorakenne, jossa käsittelyyn otetaan kerrallaan yksi taulun sarakkeen arvo. Arvo luetaan muuttujaksi ja sitä voidaan käyttää kierroksessa suoritettavien kyselyiden, funktioiden tai proseduurien yhteydessä, ja myös taulun päivitys on mahdollista jokaisella kierroksella. Tässä tapauksessa kursorina oli sata ammattinimikkeen kirjoitusmuotoa sisältänyt taulun sarake, josta käsittelyyn otettiin kerrallaan yksi. Kirjoitusmuodolle laskettiin editointietäisyys kaikkien synonyymitaulun synonyymien kanssa ja tulokset vietiin uuteen tauluun.

Testissä kävi ilmi, että yksi löytämistäni Levenstein-algoritmin toteutuksista oli noin viisi kertaa nopeampi kuin yrityksemme aiemmin käyttämä toteutus. Otin nopeimman algoritmin itselleni käyttöön ja myöhemmin sama versio algoritmista otettiin käyttöön myös muualla yrityksessä. Löytämäni nopein versio Levenstein-algoritmista on esitetty kuvassa 3.

```
CREATE FUNCTION [dbo].[udf_levenshtein](@s1 nvarchar(3999), @s2 nvarchar(3999))
RETURNS int
AS
BEGIN
  DECLARE @s1_len int, @s2_len int, @i int, @j int, @s1_char nchar, @c int, @c_temp int,
    @cv0 varbinary(8000), @cv1 varbinary(8000)
  SELECT @s1_len = LEN(@s1), @s2_len = LEN(@s2), @cv1 = 0x0000, @j = 1, @i = 1, @c = 0
  WHILE @j <= @s2_len
    SELECT @cv1 = @cv1 + CAST(@j AS binary(2)), @j = @j + 1
  WHILE @i <= @s1_len
  BEGIN
    SELECT @s1_char = SUBSTRING(@s1, @i, 1), @c = @i, @cv0 = CAST(@i AS binary(2)), @j = 1
    WHILE @j <= @s2_len
    BEGIN
      SET @c = @c + 1
      SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j-1, 2) AS int) +
        CASE WHEN @s1_char = SUBSTRING(@s2, @j, 1) THEN 0 ELSE 1 END
      IF @c > @c_temp SET @c = @c_temp
      SET @c_temp = CAST(SUBSTRING(@cv1, @j+@j+1, 2) AS int)+1
      IF @c > @c_temp SET @c = @c_temp
      SELECT @cv0 = @cv0 + CAST(@c AS binary(2)), @j = @j + 1
    END
    SELECT @cv1 = @cv0, @i = @i + 1
  END
  RETURN @c
END
```

Kuvio 3. Levenstein-funktion lähdekoodi

5.7.2 Synonyymien etsintä Levenstein-algoritmin avulla

Vertailujoukko eli erilaisten kirjoitusmuotojen määrä synonyymitaulussa oli tässä vaiheessa kasvanut jo melko suureksi (yli 70.000), joten uudesta, nopeammasta Levenstein-algoritmista huolimatta kaikkien synonyymi/syötetty ammattinimike -kombinaatioiden editointietäisyyksien laskeminen olisi ollut aivan liian raskasta, sillä vertailemalla kaikkia synonyymeja haettaviin ammattinimikkeisiin tuloksia olisi tullut yli puolitoista miljardia (jokaiselle haettavalle ammattinimikkeelle 70.000 tulosta). Päätin soveltaa käsittelyä niin, että jokaiselle haettavalle ammattinimikkeelle otettaisiin vertailujoukoksi synonyymeista vain ne, joiden kolme ensimmäistä merkkiä ovat samat kuin haettavassa ammattinimikkeessä. Koska ammattinimikkeissä nimikkeen merkitys voi olla täysin erilainen jo esimerkiksi kahden merkin vaihtuessa, päätin tarkastella aluksi vain niitä ammattinimikepareja, joiden Levenstein-etäisyys on yksi eli eroavat toisistaan vain yhden merkin osalta. Muodostuneista pareista virheelliset korjattiin niin, että oikeat synonyymit syötettiin käsin synonyymitauluun ja loput lisättiin sellaisenaan automaattisesti synonyymeiksi. Sama käsittely tehtiin seuraavaksi niin, että vertailujoukoksi otettiin synonyymeistä ne, joiden kolme viimeistä merkkiä olivat samat kuin haettavassa ammattinimikkeessä. Huomioitavaa tietysti on, että vaikka useimmiten tapauksessa, jossa vain yksi kirjain eroaa henkilön syöttämän ja virallisen kirjoitusmuodon välillä,

tarkoittavat kirjoitusmuodot samaa ammattinimikettä, myös poikkeuksia löytyy. Esimerkiksi merkkijonojen ”erottaja” ja ”verottaja” editointietäisyys on yksi, vaikka sanojen merkitys on hyvin erilainen. Näistä ”verottaja” ei ole Tilastokeskuksen ammattinimikkeistä löytyvä ammatti, mutta väestötietojärjestelmään syötettynä kyseinen kirjoitusmuoto kuitenkin löytyy kolmelta henkilöltä, joten suoraan hyväksymällä kaikki synonyymit, joiden editointietäisyys on yksi, saavat nämä kolme henkilöä viralliseksi ammattinimikkeekseen ”erottaja”.

Tämä käsittely, joka vaatii joko kolmen ensimmäisen tai kolmen viimeisen merkin olevan samat haettavien ammattinimikkeiden virallisten kirjoitusmuotojen kanssa, ei luonnollisesti ole täysin kattava etsittäessä pareja, joiden kirjoitusmuodot eroavat vain yhden merkin osalta, sillä eroavaisuus saattaisi olla myös joko henkilön ilmoittaman nimikkeen alun tai lopun kolmessa merkissä. Tavoitteena olikin, että tämän käsittelyn avulla saataisiin löytymättömien ammattinimikkeiden joukkoa pienennettyä entisestään, sillä jokainen löydetty ammattinimikepari vähentää edellä mainittua kaikkien haettavien ammattinimikkeiden ja kaikkien synonyymien vertailun tuloksia aina synonyymien määrällä eli noin 70.000 rivillä per haettava nimike.

5.8 Ammattinimikkeisiin sisältyvät ammattinimikkeet

Tarkkaillessa sitä joukkoa haettavista ammattinimikkeistä, joille ei vielä ollut löytynyt vastinetta virallisista nimikkeistä, havaitsin, että monet VTJ:ään syötetyistä ammattinimikkeistä olivat tarkempia kuin vastaava virallinen ammattinimike. Jos henkilö on syöttänyt ammatikseen esimerkiksi ”yhdyskuntatyönsihteeri”, saisi se osua viralliseen ammattinimikkeeseen ”sihteeri”. Samantapainen ongelma oli tapauksissa, joissa henkilö on syöttänyt itselleen kaksi ammattia: ”luokanopettaja/historianopettaja” saisi osua viralliseen ammattinimikkeeseen ”luokanopettaja” tai ”historianopettaja”. Tästä sain idean tehdä vielä vertailu tietyin ehdoin, joilla synonyymiksi voitaisiin suoraan lisätä haettava ammattinimike, mikäli haettava nimike sisältyisi kokonaisuudessaan synonyymiin tai toisinpäin.

5.9 Proseduuri merkkijonojen pisimmän yhteisen merkkijonon löytämiseksi

Lisäksi tein proseduurin, joka etsii kahdesta annetusta merkkijonosta pisimmän yhteisen merkkijonon. Yhteistä merkkijonoa haetaan systemaattisesti kaikista eri merkkijonon alkioista alkaen ja talteen otetaan näistä pisin yhteinen merkkijono. Suhteuttamalla saatu tulos joko haettavaan tai synonyymiammattinimikkeeseen saadaan yhteisen merkkijonon pituus haettavan merkkijonon suhteen, joka palautetaan proseduurista. Jos tämä suhde ylittää tietyn rajan, voidaan ammattinimikepari lisätä synonyymitauluun. Suhdeluvun raja on tarkoitus määritellä testausvaiheessa, jotta nähdään, millä raja-arvoilla proseduuri toimii halutulla tavalla.

5.10 Uusien synonyymien generointi koneellisesti

Paljon lisää käyttökelpoisia synonyymeja sain lisäksi generoimalla koneellisesti erilaisia kirjoitusmuotoja virallisista ammattinimikkeistä. Lisäsin esimerkiksi kaikista ammattinimikkeistä, jotka päättyvät merkkijonoon ”johtaja”, saman merkkijonon, jossa johtaja on vaihdettu päälliköksi. Tällöin esimerkiksi ammattinimike ”aikuiskoulutusjohtaja” sai synonyymin ”aikuiskoulutuspäällikkö”. Samalla logiikalla korvaamalla merkkijonon ”työntekijä” merkkijonolla ”työläinen” vanerityöntekijä saa synonyymin vanerityöläinen.

5.11 Synonyymien lisääminen käsin

Käsin lisättyjä synonyymeja lisäsin pitkin järjestelmän rakentamista sitä mukaa kun havaitsin tarvetta. Erityisen vaikeita olivat tapaukset, joissa kaksi virallista ammattinimikettä ovat kirjoitusmuodoltaan hyvin lähellä toisiaan (esimerkiksi sahaaja/vahaaja) tai joissa samaa tarkoittavat ammattinimikkeet ovat kirjoitusmuodoltaan hyvin erilaiset (esimerkiksi telemyyjä = puhelinmyyjä).

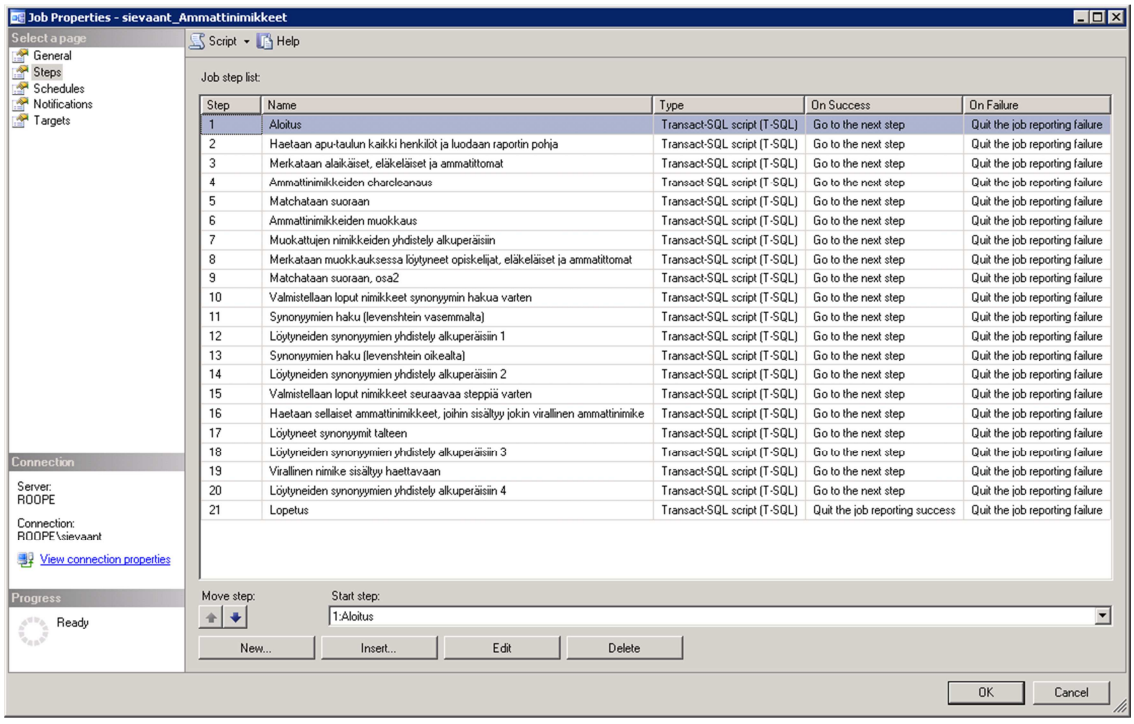
5.12 Roskakori

Niitä ammattinimikkeitä varten, joille ei löytynyt vastinetta virallisista ammattinimikkeistä, tehtiin roskakori-taulu. Tarkoituksena on, että ensimmäisten ajojen jälkeen roskako-

riin päätyneitä ammattinimikkeitä käytäisiin läpi säännöllisin väliajoin. Alun jälkeen löytymättömiä ammattinimikkeitä kertyy huomattavasti vähemmän, joten tällöin satunnainen läpikäynti riittää roskakorin pitämiseen puhtaana.

6 Järjestelmän kokoonpano ja testaus

Proseduurien ollessa valmiita alkoi järjestelmän kokoaminen. Perusrunkona järjestelmälle on SQL-kielessä suuremmassa toimintokokonaisuudessa yleisesti käytetty jobi. Jobi koostuu erinäisestä määrästä vaiheita, joihin jokaiseen voidaan tallettaa tietty toiminnallisuus. Jobia ei ole pakko suorittaa kerralla kokonaan, vaan jobi voidaan määrätä käynnistymään tietystä vaiheesta tai vastaavasti loppumaan tiettyyn vaiheeseen. Jokainen vaihe palauttaa tiedon siitä, onko vaihe suoritettu onnistuneesti vai ei ja tämän paluuarvon avulla jobin suoritusta voidaan ohjata esimerkiksi vaiheen epäonnistuessa siirtymään lähettämään sähköpostia jobin ylläpitäjälle. Valmiin jobin rakenne eri vaiheineen on esitetty kuvassa 4.



Kuvio 4. Valmiin jobin rakenne

Jobi etenee vaiheittain suunnitteluvaiheessa laatimani kaavion mukaisesti ja kattavat vaiheiden kuvaukset helpottavat järjestelmän testaamista osissa. Aluksi jobin kaikki vaiheet asetettiin lopettamaan jobi vaiheen onnistuneen suorituksen jälkeen. Tämän jälkeen oli helppo käydä tutkimassa muokattujen ja mahdollisesti syntyneiden taulujen rakenteita ja sisältöjä. Myös mahdollisten virheiden jäljittäminen on tällä tavoin helpompaa pystyttäessä suorittamaan eräänlaista haarukointia ongelmakohdan löytämiseksi.

Kokonaisuuden testaaminen oli sinänsä helppoa, sillä jobin vaiheiden koostuessa pääosin valmiiksi tehdyistä ja testatuista proseduureista oli jokaisen vaiheen sisältämä toiminnallisuus lähinnä proseduurien ajoja ja yksinkertaisia taulujen päivityksiä niiden suoritusten jälkeen, jotta tulokset saatiin vietyä tauluihin ennen seuraavaan vaiheeseen siirtymistä.

7 Tuotantoon siirto

Lukuisien testikierrosten jälkeen järjestelmä siirrettiin tuotantoon normaaleiden, sunnuntaisin ajettavien kannan päivitysajojen yhteyteen. Kun tarvittavat kannat on ensin luotu, käynnistetään ammattinimikkeiden luokittelu, jonka suoritus kestää noin 1,5 tuntia. Jokaisesta suorituskerrasta syntyy raportti, joka on esitetty kuvassa 5. Raporttiin kirjautuvat lukumäärät jokaisessa suorituksen vaiheessa seurantaa varten. Tämä helpottaa seuraamaan järjestelmän toimintaa, sillä jo raportin läpilukeminen antaa käytännössä kaikki tarvittavat tiedot järjestelmän suorituksesta, kuten alaikäisten, eläkeläisten ja ammatittomien lukumääristä, eri vaiheessa löytyneistä synonyymeistä sekä roskakoriin päätyneiden ammattinimikkeiden lukumäärästä.

Selite	Kappaletta	Päivämäärä
Kaikki henkilöt	5077331	3.4.2011 11:35
Kaikki ammatit	166357	3.4.2011 11:35
		3.4.2011 11:35
alaikäisiä	1054382	3.4.2011 11:38
eläkeläisiä	726398	3.4.2011 11:38
ammatittomia	725336	3.4.2011 11:38
jatkoon henkilöitä	2571215	3.4.2011 11:38
jatkoon ammatteja	129563	3.4.2011 11:38
		3.4.2011 11:38
Charcleanauksen jälkeen eri ammatteja	105882	3.4.2011 11:42
		3.4.2011 11:42
Charcleanauksen jälkeen osuu suoraan henkilöitä	2460866	3.4.2011 11:45
Charcleanauksen jälkeen osuu suoraan ammatteja	9026	3.4.2011 11:45
Charcleanauksen jälkeen jatkoon henkilöitä	110349	3.4.2011 11:45
Charcleanauksen jälkeen jatkoon ammatteja	34750	3.4.2011 11:45
		3.4.2011 11:45
Muokkauksen jälkeen eri ammatteja	23439	3.4.2011 11:49
		3.4.2011 11:49
Muokkauksen jälkeen löytyneitä opiskelijoita, eläkeläisiä ja ammatittomia	66522	3.4.2011 11:50
Muokkauksen jälkeen jatkoon henkilöitä	43827	3.4.2011 11:50
Muokkauksen jälkeen jatkoon ammattinimikkeitä	23749	3.4.2011 11:50
		3.4.2011 11:50
Muokkauksen jälkeen osuu suoraan henkilöitä	9907	3.4.2011 11:50
Muokkauksen jälkeen osuu suoraan ammattinimikkeitä	1638	3.4.2011 11:50
Muokkauksen jälkeen jatkoon henkilöitä	33920	3.4.2011 11:50
Muokkauksen jälkeen jatkoon ammattinimikkeitä	22111	3.4.2011 11:50
		3.4.2011 11:50
Synonyymihaun jälkeen osuu suoraan henkilöitä	5679	3.4.2011 11:58
Synonyymihaun jälkeen osuu suoraan ammatteja	2516	3.4.2011 11:58
Synonyymihaun jälkeen ei osu henkilöitä	28241	3.4.2011 11:58
Synonyymihaun jälkeen ei osu ammatteja	19595	3.4.2011 11:58

Kuvio 5. Raportti suorituksen jälkeen

8 Tulokset

Aloin seurata vanhalla systeemillä luokittelemattomien ammattinimikkeiden määrää toukokuussa 2010, jolloin luokittelemattomiin ammattinimikkeisiin kuuluvia täysi-ikäisiä henkilöitä Suomen väestöstä oli 32 127 kpl. Otin itselleni yhdeksi tavoitteeksi, että uudella järjestelmällä luokittelemattomia henkilöitä tulee olla vähemmän kuin ennen. Tärkein tavoite kuitenkin oli se, että kaikki lisättävät synonyymit olisivat oikein, sillä virheellisten synonyymien korjaaminen jälkeenpäin olisi huomattavan työlästä, sillä se tulisi tehdä käsin.

Maaliskuussa 2011 vanhalla systeemillä luokittelemattomiin ammattinimikkeisiin kuuluvien henkilöiden määrä oli kasvanut 40 903 kappaleeseen kirjoitusmuotojen lisääntyessä ja ammattinimikkeiden monipuolistuessa. Vastaavasti rakentamani järjestelmän avulla luokitelluista ammattinimikkeistä löytymättömiä henkilöitä oli 28 241 kappaletta, joten tavoitteeni paremmasta luokitteluprosentista oli jo tässä vaiheessa toteutunut. Järjestelmää olisi mahdollista ajaa myös löysemmillä asetuksilla, jolloin löytymättömien ammattien lukumäärä laskisi, mutta tähän en vielä halunnut lähteä, sillä tämä lisäisi myös väärien synonyymien mahdollisuuksien määrää. Nämä 28 241 henkilöä, joiden antamalla ammattinimikkeelle ei löydetty synonyymiä virallisista, tarkoittavat 19 595 erilaista ammattinimikkeiden kirjoitusmuotoa. Seuraavana vuorossa onkin löytymättömien nimikkeiden tutkiminen, jotta saataisiin niiden kesken yhteneväisyyksiä ja näin luotua sitä kautta lisää käsittelysääntöjä, lisättyä hankalia kirjoitusmuotoja käsin ja niin edelleen. Hyvä on pitää kuitenkin mielessä, että kaikkia väestötietojärjestelmästä löytyviä ammattinimikkeiden kirjoitusmuotoja ei ole edes mahdollista kohdistaa virallisiin ammattinimikkeisiin, sillä osa on liian virheellisiä ymmärrettäväksi edes käsin läpi käymällä.

Myös löydettyjen eläkeläisten ja opiskelijoiden määrät kasvoivat noin 470 000:sta 1 040 00:een. Vaikka nämä ryhmät eivät varsinaisesti ole arvokkaita kohderyhmien näkökulmasta, parantaa niiden löytyminen jäljelle jääneen kohderyhmän arvoa, sillä kaikki löytyneet eläkeläiset ja opiskelijat saadaan jätettyä kohderyhmän ulkopuolelle.

8.1 Validointi

Toiminnallisuuden virheettömyyden tutkiminen osoittautui vaikeaksi suorittaa koneellisesti, joten synonyymien oikeellisuutta testatakseni kävin käsin läpi 500 erilaista, satunnaisesti poimittua väestötietojärjestelmän ammattinimikkeen kirjoitusmuotoa, joille vastine oli järjestelmän avulla löytynyt. Jätin näistä pois myös kaikki tapaukset, joissa kirjoitusmuoto osui virallisiin nimikkeisiin suoralla vertailulla sekä kaikki ammatittomat, opiskelijat ja eläkeläiset. Vein nämä viisi sataa kirjoitusmuotoa löytyneen vastineensa kanssa taulukkoon ja kävin ne läpi Excelissä. Näistä neljä tapausta oli saanut parikseen selkeästi väärän ammattinimikkeen, joten virheprosentti tässä otoksessa on 0,8 prosenttia. Vaikka otos onkin erilaisten kirjoitusmuotojen suhteen melko pieni, saadaan tästä suuntaa antava arvio virheiden määrästä.

Virheelliset synonyymit olivat löytyneet nimikkeille ”kotit op”, ”lv-asentaja”, ”host” ja ”korvausneuvoja”. Näistä ”kotit op” oli saanut synonyymikseen nimikkeen ”kotiopettaja”, vaikka oikea nimike olisi ollut kotitalousopettaja. Virhe selittyy sillä, että järjestelmään ei ole erikseen määritelty lyhennettä ”kotit”, joten sen jälkeen, kun lyhenne ”op” on avattu muotoon opettaja, on muokatun merkkijonon ”kotitopettaja” ja löydetyn synonyymin, ”kotiopettaja”:n editointietäisyys yksi. Tällöin löydetty synonyymi täyttää sille vaaditun ehdon, ja synonyymi on lisätty synonyymitauluun. Sama ongelma on kyseessä kahdessa muussakin virheellisessä synonyyminimikkeessä. Nimikkeen ”lv-asentaja” tapauksessa, jossa synonyymiksi on löytynyt ammattinimike ”tv-asentaja”. Oikea vastine olisi kuitenkin ollut lvi-asentaja. Vastaavasti nimikkeen ”korvausneuvoja” synonyymiksi on valikoitunut ”korjausneuvoja” sillä vastaavanlaista virallista ammattinimikettä ei suoraan löytynyt. Viimeinen virheellinen synonyymi on löydetty nimikkeelle ”host”, joka on saanut synonyymikseen nimikkeen ”hoitaja”, sillä hoitajan lyhenteen ”hoit” ja haetun nimikkeen editointietäisyys on yksi. Tässä tapauksessa saadaan esimerkki henkilön syöttämästä nimikkeestä, joka on virheellinen tai vähintään hyvin haastava, sillä on vaikea sanoa mitä ammattia nimikkeellä ”host” ylipäätään tarkoitetaan.

Uusien muodostettujen synonyymien tarkistamisesta ja löytymättömien ammattinimikkeiden kohdentamisesta oikeaan vastapariin on tarkoitus tulla jatkuva ylläpitotyö, johon käytettäisiin mahdollisesti korkeintaan yksi henkilötyötunti per viikko uuden kannan muodostamisen ja järjestelmän ajamisen jälkeen.

8.2 Jatkokehitysideoita

Tämän laajuisessa projektissa ilmenee luonnollisesti jatkokehityksen kohteita kehitystyön eri vaiheissa. Helpoimmin toiminnallisuuteen integroitavat kehitysideat toteutin tietysti työtä tehdessä, mutta isommat jatkokehitysideat jäivät tehtäviksi tämän työn ja projektin ulkopuolella. Suurin este näiden toteuttamiseen tässä työssä oli aikataulu, jossa systeemi oli saatava yrityksessämme tuotantoon.

Jatkokehityksen kohteista yksi voisi olla jonkinlainen kirjoitusvirheiden simulointi synonyymeja luotaessa, sillä suuri osa (muuttoilmoituksen kautta tulevat sekä maistraatteihin ilmoitetut) ammattinimikkeistä skannataan paperilta sähköiseen muotoon, joten

suomen kielen yleisimpiä kirjoitusvirheitä löytyy näistä varmasti. Vaarana on tietysti ammattinimikkeiden kohdentuminen väärään synonyymiin, jonka takia huolelliseen testaukseen tulisi käyttää paljon aikaa.

Toinen, lähinnä ulkomaalaisia kirjoitusmuotoja varten tarvittava, ominaisuus voisi olla lausumiseen perustuva ”sounds like” -funktio, joka muodostaisi eri kirjoitusmuotoja sen perusteella, miltä sanan lausuminen kuulostaa. Esimerkiksi englantilainen ammattinimike ”manager” kuulostaa lausuttuna suomalaisittain jotakuinkin ”mänitser”:ilta. SQL-kielestä löytyy sisäänrakennettuina *soundex*-funktio, joka palauttaa nelimerkkisen koodin annetusta merkkijonosta, sekä *difference*-funktio. Difference-funktio palauttaa kahden annetun merkkijonon muodostamien soundex-koodien välisen eron asteikolla yhdestä neljään, jossa yksi kuvaa merkkijonojen olevan ääntämiseltään hyvin erilaiset ja neljä taas niiden olevan hyvin samanlaiset. Näiden ominaisuuksien hyöty arvioitiin kuitenkin kyseenalaisiksi, sillä suomen kielessä funktiot toimivat hyvin huonosti ja englanninkielisiä ammattinimikkeitä on suhteessa todella vähän. Suomen kieltä tukevaa versiota soundex-funktiosta en löytänyt ja sellaisen toteuttaminen vaatisi todella paljon resursseja.

Kokonaan toteuttamatta jäi ammattinimikkeelle haettujen synonyymiehdokkaiden pisteytys, sillä pisteytyksen toteutus synonyymiehdokkaille osoittautui hankalaksi rakentaa siten, että tulokset olisivat luotettavia. Tunnistusjärjestelmässä vastineita etsittäessä nimen ja osoitteen eri elementit mahdollistavat niiden pisteytyksen helpommin, jolloin esimerkiksi haettavan osoitteen ja synonyymiehdokkaan postinumeron ollessa samat voidaan kyseiselle ehdokkaalle antaa tietty määrä pisteitä. Ammattinimikkeiden tapauksessa taas elementtejä on vain yksi, joten pisteytys tulisi tehdä mahdollisesti perustuen pelkästään sanojen samankaltaisuudelle ja merkkijonojen samojen kirjainten määrään. Lisää haastetta tuo vielä mahdollisten haettavan nimikkeen ja virallisen nimikkeen kirjoitusmuotojen suuret erot esimerkiksi merkkijonojen pituuksissa, jolloin koneellisesti synonyymin oikeellisuuden päätteleminen vaikeutuu.

Toteuttamatta jäi ajanpuutteen vuoksi myös vaihe, missä löytymättömille ammattinimikkeille laskettaisiin editointietäisyys kaikkia synonyymitaulun synonyymejä vastaan. Sillä kaikkien löytymättömien ammattinimikkeiden käsittely olisi edelleen hyvin raskasta, työstän todennäköisesti tätä jatkossa oman työni ohella eräänlaisena eräajona, jossa käsittelyyn otettaan kerrallaan osa löytymättömistä.

Käytännöllinen työkalu löytymättömien ammattinimikkeiden käsittelyyn olisi tietysti graafinen käyttöliittymä, jonka näyttäisi toisessa sarakkeessa löytymättömät nimikkeet ja toisessa viralliset kirjoitusmuodot. Nämä yhdistämällä synonyymi lisättäisiin automaattisesti synonyymitauluun.

Toiminnallisuuden toteuttamisen jälkeen tutustuin itselleni uuteen työkaluun, SQL Server Reporting Serviceen, jonka avulla pystytään luomaan huomattavasti asiallisempia raportteja muodostamalla raportin pohja graafisen käyttöliittymän avulla ja käyttämällä kyselyiden tuottamaa dataa raportin täyttämisessä. Järjestelmän tuottamaa raporttia käytetään tällä hetkellä ainoastaan yrityksen sisäisessä käytössä, joten sen ulkomuoto on toisarvoinen asia, mutta tarvittaessa Reporting Servicen avulla raportista saisi varsin mallikelpoisen.

9 Yhteenveto

Työn tarkoituksena oli kehittää Bisnode Finland Oy:n käyttöön järjestelmä, joka lajittelisi väestötietojärjestelmästä yrityksen käyttöön saamat henkilöiden ilmoittamat, hyvin monimuotoiset ammattinimikkeet Tilastokeskukselta tilattuihin virallisiin ammattinimikkeisiin, jotta ammattinimikettä voitaisiin entistä paremmin hyödyntää yhtenä poimintaehtona kohderyhmiä asiakkaille rajattaessa.

Järjestelmän rakentaminen oli ensimmäinen suurempi projektini, jonka olen SQL-kielen saralla tehnyt. Tätä ennen SQL-kielellä ohjelmointi osaltani on käsittänyt pienempien funktioiden ja proseduurien tekemistä, joita olen luonut helpottamaan jokapäiväisiä työtehtäviä. Tästä johtuen toteutuksen suunnittelu- ja kehitystyöhön kuluvasta ajasta ei itselläni ollut projektin alussa tarkkaa kuvaa ja aikaa projektiin kuluikin huomattavasti enemmän kuin mitä olin kuvitellut. Projektin toteuttaminen oman työn ohella hankaloitti työn toteuttamista, sillä kerralla projektin tekoon varattavat ajankohdat olivat hyvin hajanaisia. Loppuvaiheessa omaa työkuormaa kevennettiin ja tätä kautta aikaa projektin toteuttamiseen varattiin enemmän, joka mahdollisti tuotantoon siirron lähes aikataulussa.

Työn huolellinen suunnittelu ennen toteutuksen aloittamista helpotti työn tekemistä huomattavasti. Aikaa suunnitteluun uhrasin paljon, mutta sen ansiosta käytännössä mitään ei tarvinnut tehdä uudelleen.

Tulosten läpikäynnin valossa voi sanoa, että olen hyvin tyytyväinen toteutukseen. Pientä paranneltavaa on jo järjestelmän valmistumisen jälkeen ilmennyt, mutta näiden parannusten toteutus on helppoa, kun järjestelmä on kaikin puolin itselle tuttu.

Tuotantoon siirron jälkeen asiakkaille, jotka ovat ennen käyttäneet vanhaa ammattiluokittelua kohderyhmien rajaamiseen, on jo tehty tarjouksia käyttää vastaisuudessa uutta, Tilastokeskuksen ammattinimikkeisiin ja -luokkiin perustuvaa luokittelua. Mitään uutta erillistä tuotetta yrityksellemme ei tästä aiota luoda, vaan tarkoituksena on tarjota asiakkaalle ammattinimikkeiden käyttö entistä parempana poimintaperusteena kohderyhmiä kasattaessa. Tätä tukevat pienentynyt tuntemattomien ammattinimikkeiden joukko sekä vähentynyt virheiden määrä ammattinimikkeissä.

Samankaltaista merkkijonojen kohdentamista on tarvetta tehdä muissakin yrityksemme kokonaisuuksissa ja palveluissa. Periaatteessa ongelmaan törmätään hyvin usein tapauksissa, joissa eri lähteistä käyttöön saatuja aineistoja on tarpeen yhdistellä keskenään voidakseen tuoda aineistoon lisää tietoa ja sitä kautta arvoa eli niin sanotusti rikastaa dataa, joka taas on yksi yrityksemme tärkeistä tekemisistä.

Jo nyt on tullut ilmi tarve tehdä samanlaista kohdentamista yritysnimikkeille, joiden täsmääminen keskenään on hyvin haasteellista johtuen niin ikään hyvin erilaisista kirjoitusmuodoista riippuen lähteestä. Käsittelysääntöjen osalta yritysnimet ovat kuitenkin täysin erilainen kohdejoukko lyhenteineen, yhtiömuotoineen ja niiden lyhennyksineen, sisaryrityksineen ja aputoiminimineen. Perustoiminnallisuus joka tapauksessa kävisi samalla tavalla näiden kohdentamiseen, sillä perustaltaan kyse on myös merkkijonojen vertailusta apufunktioiden ja käsittelysääntöjen avulla.

Lähteet

1 Henkilötietolaki, 4. luku, 19§: Suoramarkkinointi ja muut osoitteelliset lähetykset. Verkkodokumentti. <http://www.finlex.fi/fi/laki/ajantasa/1999/19990523> 22.4.1999. Viitattu 2.4.2011.

2 Väestörekisterikeskuksen Tarkasta tietosi! -palvelu. Verkkodokumentti. <https://verkkopalvelu.vrk.fi/Omat/Etusivu.aspx> Viitattu 2.4.2011.

3 Väestötietoasetus, 2. luku, 6§: Henkilötietojen ilmoittaminen ja ylläpito. Verkkodokumentti. <http://www.finlex.fi/fi/laki/alkup/1993/19930886> 22.10.1993. Viitattu 2.4.2011.

4 Suomessa asuvat lääkärit 1.1.2011. Verkkodokumentti. <http://www.laakariliitto.fi/tilastot/laakaritilastot/taskutilasto.html> 25.2.2011. Viitattu 2.4.2011.

5 Kuluttajansuojalaki, 2. luku, 2§: Markkinoinnin hyvän tavan vastaisuus. Verkkodokumentti. <http://www.finlex.fi/fi/laki/ajantasa/1978/19780038> 20.1.1978. Viitattu 2.4.2011.

6 Luettelo yliopistotutkintojen lyhenteistä. Verkkodokumentti. <http://fi.wikipedia.org/wiki/Luettelo_yliopistotutkintojen_lyhenteist%C3%A4> 17.1.2011. Viitattu 2.4.2011.

7 Levenshtein distance. Verkkodokumentti. <http://en.wikipedia.org/wiki/Levenshtein_distance > 27.3.2011. Viitattu 2.4.2011.

8 String Searching using Graph Drawing and Geometric Techniques. Verkkodokumentti. <http://www.facweb.iitkgp.ernet.in/~pabitra/facad/06CS6001t.pdf> Viitattu 2.4.2011.