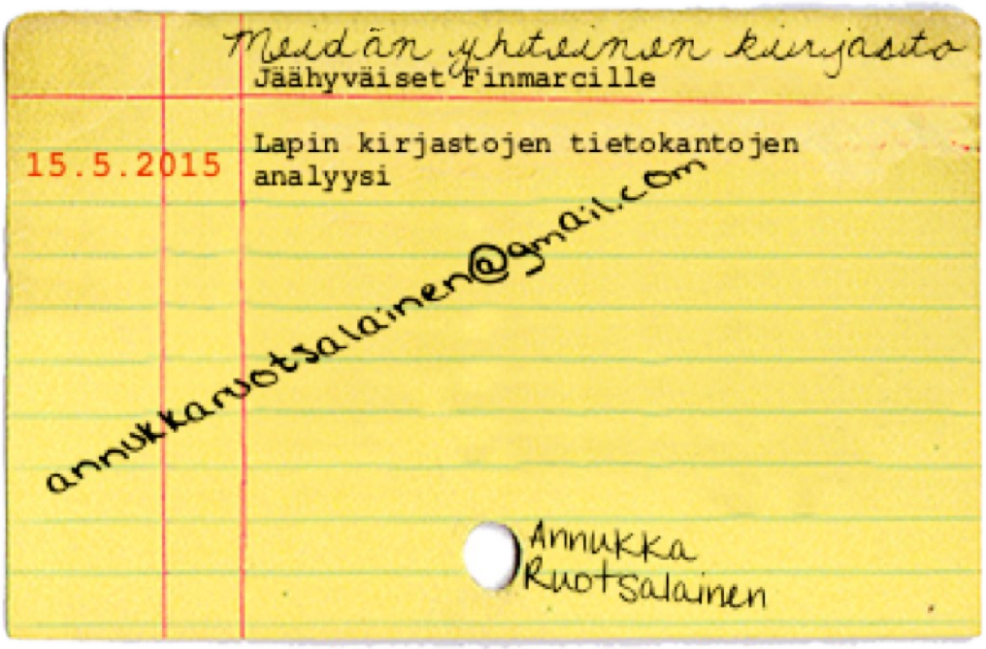
Meidän yhteinen kirjasto
Jäähyväiset Finmarcille
15.5.2015
Lapin kirjastojen tietokantojen analyysi
annukkaruotsalainen@gmail.com
Annukka Ruotsalainen

Annukka Ruotsalainen
annukkaruotsalainen@gmail.com

Jäähyväiset Finmarcille

Lapin kirjastojen tietokantojen analyysi

toukokuu 2015

Rovaniemen kaupunginkirjasto – Lapin maakuntakirjasto

Sisällys

Lukijalle

Tässä raportissa kerron Lapin kirjastojen tietokantojen analyysin tuloksista. Analyysin kohteena oli kaikkiaan viisi tietokantaa: Kemin, Keminmaan ja Posion kirjastojen sekä Lapin ja Karpalo-kirjastokimpan tietokannat.

Työ on osa Meidän yhteinen kirjasto -projektia. Lapin kirjasto-kimppaan kuuluu suurin osa lappilaisista yleisistä kirjastoista sekä joukko museo- ja oppilaitoskirjastoja. Kimppa toimii 15 kunnan alueella, ja sillä on yhteinen aineisto- ja asiakasrekisteri. Etelä-Lapin kuuden kunnan (Kemi, Keminmaa, Posio, Simo, Ranua, Tervola) kirjastot ovat tehneet periaatepäätöksen liittymisestä Lapin kirjasto-kimppaan tulevassa järjestelmänvaihdossa.

Lapin kirjastossa on käytössä Axiell oy:n PallasPro-kirjastojärjestelmä, Etelä-Lapin kirjastoissa puolestaan saman firman Origo. Kemin ja Keminmaan kirjastot eivät kuulu mihinkään kimppaan. Posion kirjastolla on yhteinen verkkokirjasto Kuusamon kirjaston kanssa. Ranua, Simo ja Tervola ovat osa Karpalo-kirjastokimppaa.

Yleisissä kirjastoissa on käytössä Finmarc-luettelointiformaatti, joka on vaihtumassa uuteen, Marc 21 -formaattiin. Tämä edellyttää myös kirjastojärjestelmän vaihtoa uutta formaattia käyttävään järjestelmään. Pohjoisen Suomen kirjastoissa uusi luettelointiformaatti ja kirjastojärjestelmä on päätetty ottaa käyttöön vuoden 2016 aikana.

Suositeltavaa on, että tietokanta siivotaan eli bibliografiset tietueet tarkistetaan ennen tietokannan konversiota. Konversio ei koskaan onnistu täysin oikein, mutta ennen konversiota tehtävällä siivoamisella vältetään ainakin osa ongelmista. Jotta siivoaminen ei olisi täysin mielivaltaista, on hyvä tutkia, millaisia virheitä tietokannassa on ja päättää, miten virheistä päästään eroon. Tietokantaa voi analysoida testikonversion avulla ja tutkimalla tietokannan datakopiota eli dumppia eri välinein.

Vastaavaa tietokannan analyysiä on tehty ainakin Tampereen kaupunginkirjastossa. Mace Ojalan tekemä PIKI-tietokannan siivousmenetelmät -raportti on ollut lähteenä omassakin työssäni. Keskeinen ero oli kuitenkin se, että analysoimani bibliografinen data oli Finmarc-formaatin mukaista, kun taas Tampereen tietokanta oli jo konvertoitu Marc 21 -muotoon.

Tekemäni analyysin tavoitteena oli tuottaa kirjastojen käyttöön konkreettisia listauksia virheistä, joita tietokannat sisältävät. Näiden listojen perusteella kirjastoilla on mahdollisuus päättää miten tietokantaa siivotaan ennen konversiota ja ylipäätään nähdä, millaisessa kunnossa tietokanta on.

Työn kuvaus

Maaliskuun puolivälistä toukokuun puoliväliin analysoin kirjastojen bibliografista dataa XQuery-kyselykieltä käyttäen. Data oli XML-muodossa. Keskeisin työväline oli BaseX, joka on netistä vapaasti saatavilla oleva, avoimen lähdekoodin ohjelmisto. Tein työn omalla kannettavalla tietokoneellaani etätyönä.

Tarkoituksena oli keskittyä sellaisiin virheisiin, joita kirjastojärjestelmällä ei suoraan pysty hakemaan. Etsin datasta erilaisia virheellisiä arvoja, puuttuvia kenttiä, kenttien toistumia ja muita virheitä, jotka voivat aiheuttaa ongelmia konversiossa.

Lapin kirjastolla ei ollut itse mahdollista ottaa PallasPro-järjestelmästä tietokantadumppia, vaan se täytyi tilata järjestelmätoimittajalta. Dumppi ladattiin maaliskuussa. Muissa kirjastoissa tietokantadumpin kukin kirjasto latasi itse Origosta.

Indeksoin data-aineiston BaseX-tietokantaohjelmistolla. Lisäksi tein tietokannoille testikonversion Usemarcon-ohjelmalla ja katsoin millaisia virheitä ohjelman virheloki tuottaa. Tein myös jokaisesta tietokannasta listat, mitä luettelointikenttiä kirjastot ovat käyttäneet ja kuinka paljon.

Ennen työn alkua pidettiin yksi palaveri Rovaniemellä, jossa käytiin läpi kirjastojen toiveita työn suhteen. Lisäksi pidettiin yksi etäkokous huhtikuun lopussa, jossa käytiin läpi tuottamiani virhelistoja. Muuten yhteydenpito sujui sähköpostin ja Google Driven kautta, jonne latasin virhelistoja.

Tiettueiden määrän perusteella tietokannat näyttivät tältä:

1. Tietokantojen koot (osakohteet tietueita, joissa kenttä 773)

tietokanta	tietuemäärä	osakohteiden määrä
Lappi	1244237	903991
Kemi	485579	347848
Keminmaa	167099	99479
Posio	121939	68066
Karpalo	368384	208000

Testikonversio Usemarcon-ohjelmalla

Usemarcon-ohjelman avulla voi testata, miten konversio uuteen luettelointiformaattiin onnistuu. Ohjelma on saatavissa Kansalliskirjaston sivuilta.

Ohjelma tuottaa virhelokin, josta käyvät ilmi kaikki konversiossa havaitut tietueiden virheellisyydet. Konversiosäännöt, joiden mukaan ohjelma tietueita tarkistaa ja kenttiä yhdistää, on tehty Kansalliskirjastossa. Näihin sääntöihin kirjastolla on mahdollisuus tehdä muutoksia tai lisäyksiä tarpeidensa mukaan.

Mitä Usemarcon siis tekee? Se

- tarkistaa sisään tulevan Finmarc-muotoisen datan ja vertaa sitä formaattikohtaisiin sääntöihin
- tekee muunnoksen annettujen konversiosääntöjen mukaisesti Finmarcista Marc 21 -muotoon
- tarkistaa ulos menevän Marc21 -muotoisen datan ja vertaa sitä formaattikohtaisiin sääntöihin.

Taulukossa 2 on esitetty Usemarconin antamat virheet Finmarc-muotoisessa datassa ja niiden esiintymät tietokannoissa. Lukumäärä ei tarkoita tietueiden lukumäärää, vaan virheiden esiintymiä. Virhe kun voi esiintyä yhdessä tietueessa useampaan kertaan.

2. Usemarconin antamat virheet Finmarc-datassa

Virhetyyppi / lukumäärä tietokannassa	Lappi	Kemi	Keminmaa	Posio	Karpalo
Tietueessa käytetty formaatin ulkopuolista kenttää	20327	51	3	2	257
Tietueesta puuttuu pakollinen kenttä	32098	8731	1901	236	2460
Tietueesta puuttuu pakollinen osakenttä	32929	6368	1719	14207	27838
Tietueessa väärä ensimmäinen indikaattori	1016	1008	231	412	3058
Tietueessa väärä toinen indikaattori	44076	20863	25816	4963	22693
Tietueessa väärä/ylimääräinen osakenttä	38843	242770	209591	2231	18374
Tietueessa toistettu kenttää, jota ei saa toistaa	2706	1305	1352	447	1769
Tietueessa tuntematon merkki	213505	-	-	47	-
Tietueessa kenttä, jolla ei ole osakenttää	5	14	-	2	55

Kaikkiaan tietokannoissa oli tietueita, joissa Usemarconin mukaan oli yksi tai useampi virhe, seuraavasti:

3. Virheellisten tietueiden lukumäärät Finmarc-datassa

tietokanta	virheellisiä tietueita	%-osuus koko tietokannasta
Lappi	137110	11,0
Kemi	130812	26,9
Keminmaa	42737	25,6
Posio	22547	18,5
Karpalo	76585	20,8

Millaisia virheet sitten olivat? Lapin tietokannassa formaatin ulkopuoliset kentät ovat todennäköisesti seurausta kopioluetteloinnista esimerkiksi ulkomaisesta tietokannasta. Muissa tietokannoissa formaatin ulkopuoliset kentät olivat yksittäisiä esiintymiä.

Pakollisen kentän puuttuminen tarkoitti kaikissa tietokannoissa nimekekentän 245 puuttumista. Hyvin suuri osa näistä tietueista oli osakohteita.

Pakollisten osakenttien puuttuminen koski useita eri kenttiä, mutta kaikissa tietokannoissa kenttään 773 liittyvät osakenttien puuttumiset olivat varsin yleisiä, esimerkiksi Lapissa 80 prosenttia virheistä liittyi tähän kenttään.

Indikaattoreiden virheellisyydet olivat tapauksia, jossa oli käytetty indikaattoria sellaisen kentän yhteydessä, jossa sitä ei saa käyttää tai oli käytetty väärää indikaattoria.

Väärä/ylimääräinen osakenttä tarkoitti sellaisen osakentän toistamista, jota ei saa toistaa tai väärän osakenttäkoodin käyttämistä. Toistettu kenttä taas tarkoitti, että kenttä, jota formaatin mukaan ei saa toistaa, oli tietueessa useammin kuin kerran.

Kenttä ilman osakenttää ei ollut lukumääräisesti iso virhe. Se sisälsi tietueita, joissa osakenttäkoodi oli virheellinen (esimerkiksi numero kirjaimen sijaan).

Tietueessa tuntematon merkki -virhe ilmoitti sellaisista merkeistä, jotka eivät ole Usemarconin merkkimuunnostaulukossa. Lapin tietokannassa tällainen merkki esiintyi kentässä 999.

Testikonversion jälkeen virheloki antoi virheitä seuraavasti:

4. Usemarconin antamat virheet Marc 21 -datassa

Virhetyyppi / lukumäärä tietokannassa	Lappi	Kemi	Keminmaa	Posio	Karpalo
Tietueesta puuttuu pakollinen kenttä	33711	8755	1903	22115	36167
Tietueesta puuttuu pakollinen osakenttä	32809	2521	2701	239	2725
Tietueessa väärä ensimmäinen indikaattori	503	8296	2729	10039	5341
Tietueessa väärä toinen indikaattori	1000	153	78	29	344
Tietueessa väärä/ylimääräinen osakenttä	17470	14012	3176	223	12469
Tietueessa toistettu kenttää, jota ei saa toistaa	245	187	35	38	658
Tietueessa kenttä, jolla ei ole osakenttää	2	8	-	-	52

5. Virheellisiä tietueita testikonversion jälkeen

tietokanta	virheellisiä tietueita
Lappi	57688
Kemi	25249
Keminmaa	3808
Posio	8609
Karpalo	54624

Kuten osin pienemmästä virheiden lukumäärästä ja virheellisten tietueiden lukumäärästä näkee, konversiossa korjaantuu joitakin virheitä. Esimerkiksi indikaattoreiden virheet ovat jälkimmäisessä taulukossa paikoin selvästi vähäisempiä. Formaatin ulkopuoliset kentät taas tippuvat muunnoksen yhteydessä automaattisesti pois, jos niitä ei erikseen haluta säilyttää.

Siinä, mitkä kentät ja osakentät ovat pakollisia tai toistettavia, on formaateissa eroja. Esimerkiksi Finmarcissa vain nimekekenttä 245 on pakollinen, mutta Marc 21:ssä nimekekentän lisäksi kenttä 008 on pakollinen.

Testikonversiossa jouduin hieman oikomaan virheitä. Tuntematon merkki, josta ohjelma ilmoitti Finmarc-datassa, on poistettu tietokannasta ennen formaatin muunnosta, koska muuten konversio ei olisi onnistunut. Asiaan on kuitenkin syytä kiinnittää huomiota, ja siivota merkki pois ennen varsinaista konversiota.

Muu analyysi

Tutkin bibliografisesta datasta lähinnä sellaisia kenttiä, joissa arvon täytyy olla tietyn muotoinen. Esimerkiksi luokkakentässä 098 pitäisi olla luokkanumeroita, kielikoodikentässä 041 kolmikirjaiminen kielikoodi, 021-kentässä tietynmittainen ISBN-koodi.

Lisäksi etsin tietueita, joista puuttuu tieto nimekkeestä kokonaan, informaatiokoodeja sisältävä 008-kenttä, minitietueita, tietueita, joista puuttuu luokkakenttä 098. Katsoin myös millaisia aineiston erityismääreitä kenttään 245 $z on tallennettu. Virheellisiä julkaisuaikoja etsin kentästä 260 $c ja $g.

Kirjastokohtaisiin kenttiin kiinnitiin myös huomiota. Listasin kirjastokohtaisten huomautuskenttien (kentät 565–599) ja kirjastokohtaisten sijaintitietokenttien (kentät 950–969) sisältöjä. Vertailin myös kirjastokohtaisten asiasanakenttien arvoja (kentät 692–698) Yleiseen suomalaiseen asiasanastoon eli katsoin kuinka paljon kirjastokohtaisissa asiasanakentissä on sellaisia asiasanoja, jotka löytyvät myös YSA:sta.

Vertailin kontrolloitujen asiasanakenttien arvoja (kentät 652–654) asiasanastoihin (YSA, Musa, Kaunokki) ja katsoin, kuinka paljon kentissä on termejä, joita ei sanastossa ole.

Puuttuva 008-kenttä

Kaikissa tietokannoissa oli tietueita, joista puuttui 008-kenttä. Kentän puuttuminen voi aiheuttaa vaikeuksia konversiossa, koska esimerkiksi tuplatietueiden yhdistäminen ei

välttämättä onnistu. Marc 21 -formaatissa kenttä on lisäksi pakollinen.

6. Puuttuva 008-kenttä

tietokanta	tietueita	suhteellinen %-osuus
Lappi	2718	0,22
Kemi	48	0,01
Keminmaa	12	0,007
Posio	45455	37,3
Karpalo	79011	21,4

Posion ja Karpalon kohdalla suhteelliset osuudet ovat melko suuria. Kenttä puuttui pääosin aineistosta, joka oli ennen vuotta 2000 ilmestynyttä. Lapin ja Posion tietokannoissa oli molemmissa lisäksi yksi tietue, jossa 008-kenttä oli kaksi kertaa.

Puuttuvat kentät 240 ja 245

Kaikista tietokannoista löytyi tietueita, joista puuttuu nimeketieto. Kentän 245 puuttumisia oli lukumääräisesti paljon (katso taulukko 2).

Koska tieto nimekkeestä on voitu merkitä yhtenäistetyn nimekkeen kenttään 240, etsin tietueita, joista puuttuvat molemmat kentät 240 ja 245. Hyvin suuri osa tällaisista tietueista oli osakohteita. Nimeketieto puuttui niistä kokonaan tai se oli merkitty väärään kenttään. Nimeketieto saattoi olla esimerkiksi pääkirjauskentässä.

7. Puuttuvat kentät 240 ja 245

tietokanta	tietueita	suhteellinen %-osuus
Lappi	741	0,06
Kemi	528	0,11
Keminmaa	29	0,02
Posio	101	0,08
Karpalo	644	0,17

Virheelliset ISBN-koodit

Kenttään 021 $a merkitään kirjan ISBN-koodi, joka on 10 tai 13 merkkiä pitkä merkkijono. Kirjaston tietokannoissa ISBN on merkitty vaihtelevasti väliviivojen kanssa tai ilman niitä.

Tekemässäni kyselyssä poistin mahdolliset väliviivat ja etsin koodeja, jotka ovat alle 10 tai yli 13 merkkiä pitkiä tai sisältävät muuta kuin numeroita tai kirjaimen X tai x.

Virheellisiä koodeja löytyi kaikista tietokannoista. Kenttään oli tallennettu paljon koodeja, jotka eivät ole virallisia ISBN-koodeja, vaan esimerkiksi äänitteen tunnuskoodi, jonka tallentamista varten on oma kenttänsä. Samaan kenttään oli myös tallennettu useampi kuin yksi ISBN. Kentässä saattoi olla pelkästään muutakin kuin sallittuja merkkejä (esimerkiksi *sid.*, *Cd*).

ISBN-koodia käytetään konversiossa tietueiden yhdistämiseen. Jos koodi on jollain tavalla virheellinen, ei yhdistäminen välttämättä onnistu ja syntyy tuplatietueita.

8. Virheelliset ISBN-koodit

tietokanta	tietueita	suhteellinen %-osuus
Lappi	702	0,3
Kemi	126	0,14
Keminmaa	156	0,14
Posio	1159	2,9
Karpalo	3726	3,7

Minitietueet

Minitietueet ovat hyvin niukoin tiedoin luetteloituja tietueita, jotka aiheuttavat päänvaivaa paitsi konversiossa myös tiedonhaussa.

Minitietueiden etsintä oli eräänlaista hakuammuntaa. Päätin, että tietue on jossain määrin puutteellinen tai keskeneräinen, mikäli siinä on vaihtuvamittaisia kenttiä 3 tai vähemmän. Kenttien määrän rajaaminen juuri tähän määrään oli täysin mielivaltaista. Jätin kyselystä pois osakohteet.

Minitietueita oli kaikissa tietokannoissa, mutta osa niistä oli ainakin jossain määrin oikeellisia tietueita. Esimerkiksi kirjastoissa lainattavissa olevat esineet on usein luetteloitu hyvin niukoilla tiedoilla. Toisaalta tällä haulla sai esiin myös syystä tai toisesta luettelointitiedoiltaan puutteellisiksi jääneitä tietueita, kuten saapumattomia hankintoja.

Lapin tietokannasta rajasin pois muun muassa lehdet, kartat ja esineet, koska näistä aineistolajeista oli merkintä 999-kentän $d-osakentässä. Kyseisissä aineistolajeissa kolmen kentän pituiset tietueet olivat jokseenkin säännönmukaisia. Origo-kirjastojen datasta en yhtä säännöllistä merkintää aineistolajista löytänyt, joten listaukseen tuli mukaan myös jonkin verran oikeellisia tietueita.

Suhteellinen osuus on laskettu koko tietokannan tietuemäärästä ilman osakohteita.

9. Minitietueiden määrät

tietokanta	tietueita	suhteellinen %-osuus
Lappi	130	0,04
Kemi	43	0,03
Keminmaa	17	0,03
Posio	7263	13,5
Karpalo	15148	9,4

Tuplatietueet

Tuplatietueiden etsintä on tunnetusti työlästä puuhaa. Niitä on varmasti enemmän tai vähemmän ainakin sellaisissa tietokannoissa, joissa on yhdistetty useampien kirjastojen bibliografista dataa. Konversiossa käytetään muun muassa ISBN-koodia tai muuta yksilöivää tunnistetta tietueden yhdistämiseen. Jos tällainen tunniste puuttuu, eivät tietueet yhdisty.

Tuplatietueiden etsintään olisi saatavilla erilaisia automaattisia vaihtoehtoja (esimerkiksi RecordManager, MarcXimiL tai MarcEditin deduplikointityökalu). Näihin en kuitenkaan ehtinyt tämän projektin puitteissa tutustua sen tarkemmin, joten tuplatietueiden metsästäminen jäi puolitiehen.

Tein kuitenkin kirjastoille tutkittavaksi taulukon, jossa on listattu tietokannan kaikista tietueista (lukuun ottamatta osakohteita) tietuetunniste, mahdolliset id-koodit (ISBN[1], ISSN[2], ISRC[3], NBN[4]), nimeke ja julkaisuvuosi. Taulukon yksi rivi vastaa siis yhtä tietokannan tietuetta ja sarake luettelointikentän arvoa. Taulukko on järjestetty nimekkeen mukaan aakkosjärjestykseen.

Taulukko vaatii manuaalista läpikäyntiä, mutta ainakin se on miellyttävämpi keino kuin tutkia tietueita kirjastojärjestelmän kautta. Taulukko on silmäiltävissä läpi suhteellisen vaivattomasti ilman tietueiden avaamista.

Virheelliset kielikoodit

Teoksen kieli merkitään kenttään 041 ($a, $b, $c). Kolmikirjaimisena koodina Finmarcissa käytetään ISO 639-2/B:n mukaista koodia, pienin kirjaimin merkittynä. Mikäli teoksessa on käytetty useita kieliä, koodit kirjoitetaan peräkkäin osakenttään.

Virheellisiä arvoja kentässä tuotti hyvin monenlaiset tavat ilmoittaa teoksen kieli: koodit oli osakentässä erotettu välilyönnillä ja välimerkillä, käytetty kielen suorasanaista muotoa

1 International Standard Book Number
2 International Standard Serial Number
3 International Standard Recording Code
4 National Bibliography Number

(esimerkiksi *suomi*, *englanti*), kolmikirjaiminen koodi tai koodien yhdistelmä oli väärä tai kentässä oli muuta siihen kuulumatonta tietoa.

Kielikoodien olisi hyvä olla standardin mukaisia, koska konversiossa 008-kentän kielikoodi otetaan Marc 21 -formaattiin 041-kentästä. Finmarcissa kun 008-kentässä ei ole merkintää kuvailtavan kohteen kielestä.

Suhteellinen osuuden laskin vertaamalla virheellisten arvojen yksittäisten esiintymien lukumäärää yksittäisten arvojen kokonaismäärään.

10. Virheelliset kielikoodit

tietokanta	virheellisiä arvoja	suhteellinen %-osuus
Lappi	418	14,2
Kemi	208	20,2
Keminmaa	143	27,7
Posio	59	17,6
Karpalo	335	27,4

Virheelliset luokkanumerot

Yleisissä kirjastoissa käytössä olevan Yleisen kymmenluokituksen luokat tallennetaan Finmarcissa 098-kenttään. YKL on saatavissa avoimena datana XML-muodossa, joten virheellisten luokkamerkintöjen etsiminen oli suhteellisen vaivatonta.

Selvitin, kuinka moni kenttään tallennettu arvo todella on YKL:n mukainen. Pyrin huomioimaan haussa myös muotoluokat.

Luokitusta päivitetään aika ajoin ja luokkia poistetaan ja uusia tulee tilalle, joten nyt virheelliset luokkanumerot ovat voineet olla aikoinaan oikeellisia. Virheitä aiheutti muun muassa se, että yhteen kenttään oli tallennettu useampi luokkanumero, luokkanumeron lisäksi kentässä oli muitakin merkkejä tai luokkanumerossa oli käytetty pilkkua pisteen sijaan. Myös luokkakenttään oli eksynyt aivan muuta kuin siihen kuuluvaa tietoa.

Suhteellinen osuuden laskin vertaamalla virheellisten arvojen yksittäisten esiintymien lukumäärää yksittäisten arvojen kokonaismäärään.

11. Virheelliset luokkanumerot

tietokanta	virheellisiä arvoja	suhteellinen %-osuus
Lappi	2623	32,2
Kemi	203	5,2
Keminmaa	1870	43,9
Posio	89	4,6
Karpalo	906	22,8

Vuosilukumerkinnät

Vuosilukuja tutkin julkaisutietokentässä 260 $c ja $g. Vuosiluvun pitäisi olla pääsääntöisesti neljä numeroa pitkä, mutta esimerkiksi hakasulkeet, cop- ja p-merkinnät vuosiluvun yhteydessä ovat nykyään ilmeisesti sallittuja.

Koska en löytänyt listaa sallituista merkintätavoista, jäi virheellisten julkaisuaikojen etsintä hieman vajaaksi. Haussa rajattiin pois kaikki yli neljä merkkiä pitkät arvot.

Listan perusteella kirjastot voivat siivota ainakin kaikkein räikeimmät virheellisyydet. Vuosiluvun paikalla saattoi olla myös esimerkiksi *1985 [Helsinki: Polarvox,1985*, *Oulun alue,1984* tai *Lasten oma kirjakerho, 1985*.

Puuttuvat luokkanumerot

Luokkamerkintä ei ole pakollinen, mutta monissa tapauksissa ainakin tärkeä hakukriteeri.

Tutkin kuinka paljon tietokannoissa on tietueita, joista puuttuu 098-kenttä. Rajasin haussa pois osakohteet eli sellaiset tietueet, joissa on kenttä 773. Hausta olisi pitänyt rajata pois kenties myös aineistolajeja, joissa luokkamerkintää ei käytetä.

Suhteellinen osuuden laskin vertaamalla puuttuvan kentän tietueita tietueiden kokonaismäärään, josta oli vähennetty osakohteiden määrä.

12. Puuttuva luokkakenttä 098

tietokanta	tietueita	suhteellinen %-osuus
Lappi	17917	5,3
Kemi	278	0,2
Keminmaa	71	0,05
Posio	7206	13,7
Karpalo	23056	14,4

Ainakin Karpalo-tietokannan suurta luokattomien tietueiden määrää selittänee se, että kirjastossa on takavuosina luetteloitu kaunokirjallisuutta ilman luokkamerkintää eikä kaikkiin tietueisiin ole jälkikäteen tietoa lisätty.

Aineiston yleismääreet

245z-kenttään kirjataan aineiston yleismäärе (esimerkiksi kirja, äänite, esine, nuotti). Siihen, millainen yleismääreen pitäisi olla, ei tietääkseni ole ohjeistusta. Mutta ainakin niiden olisi hyvä olla yhtenäiset koko tietokannassa.

Eri tietokantojen välillä ja myös tietokantojen sisällä yleismääreitä on käytetty hyvin kirjavasti. Kirjaston tehtäväksi jää päättää, onko aineiston määre esimerkiksi nuotti vai nuottijulkaisu, äänite vai äänitallenne, käytetäänkö määreen ympärillä hakasulkeita vai ei.

Puhtaasti kirjoitusvirheistä johtuvia kirjavuuksia oli myös paljon (esimerkiksi *Elektoninen aineisto*, *Videotaallenne*, *Nuottijilkaisu*).

Nämä eivät sinällään vaikuta konversion onnistumiseen, mutta vaikuttavat tietokannan siisteyteen.

13. Aineiston yleismääreet

tietokanta	245z-kentässä eri arvoja
Lappi	274
Kemi	173
Keminmaa	77
Posio	30
Karpalo	126

Kirjastokohtaiset kentät

Kirjastokohtaisista huomautuskentistä ja sijoituspaikkatiedoista koostin listat, joista näkyy, millaista tietoa kenttiin on tallennettu.

Näiden listojen avulla kirjastot voivat päättää, onko kenttien sisältämä tieto tarpeellista siirtää uuteen formaattiin vai voiko sen hävittää konversion yhteydessä. Marc21 -formaatissa kirjastokohtaisia kenttiä on käytössä paljon vähemmän kuin Finmarcissa, joten suositeltavaa on, että kirjastot mahdollisuuksien mukaan luopuisivat kentistä.

Kirjastokohtaiset kentät olivat vaihtelevasti käytössä kirjastoissa. Kaikissa niitä oli kuitenkin jossain määrin käytetty. Lapin kirjastossa kenttiä oli käytössä eniten, syynä tähän kirjastojen suuri määrä.

Vertasin kirjastokohtaisia asiasanoja Yleisen suomalaisen asiasanaston termeihin. Listasin kunkin kentän yksittäiset termit ja selvitin, kuinka moni näistä on myös YSA:ssa. Otin huomioon vain osakenttien $a, $b ja $x arvot. Suhteutin YSA:sta löytyvien asiasanojen määrän yksittäisten arvojen kokonaismäärään.

Kirjastokohtaisia asiasanoja oli käytetty kaikissa kirjastoissa, ja niistä suhteellisen iso osa on myös kontrolloidussa asiasanastossa.

14. Kirjastokohtaiset asiasanat ja YSA

tietokanta	asiasanat, jotka YSA:ssa	suhteellinen %-osuus
Lappi	2813	40,2
Kemi	268	51,2
Keminmaa	74	60,7
Posio	3573	42,3
Karpalo	2372	68,1

YSA:n, Musan ja Kaunokin asiasanat

Vertasin kontrolloitujen asiasanakenttien (YSA-kenttä 652, Musa-kenttä 653 ja Kaunokki-kenttä 654) arvoja eri asiasanastoihin eli katsoin moniko käytössä oleva termi ei todellisuudessa ole kontrolloidusta asiasanastosta. Otin huomioon osakenttien $a, $b ja $x arvot.

Musan ja Kaunokin kenttiin on sallittua tallentaa myös YSA:ssa olevia asiasanoja. Siksi vertasin termejä, jotka eivät löydy erikoissanastosta, myös YSA:an.

Asiasanoissa virheitä tuottivat esimerkiksi yhteen osakenttään kaksoispisteellä niputetut ketjutetut asiasanat. Myös teoksen nimen tallentaminen asiasanakenttään oli varsin yleistä. Esimerkiksi Musan kenttään oli tallennettu hyvin paljon musiikkikappaleiden nimiä. Virheitä tuottivat tietysti myös kirjoitusvirheet.

Toisaalta virheellisten asiasanojen määrä on tässä liian suuri, koska vertailu ei ottanut huomioon, että YSA:n ja Musan asisanoina voi käyttää myös sanastoon kuulumattomia termejä (niin sanottu vapaa indeksointi) kuten erisnimiä ja numeerisia ajanmääreitä. Ja sanastotkin muuttuvat: aikoinaan asiasanastossa ollut termi ei siellä ehkä enää olekaan.

Listasin yksittäiset asiasanojen esiintymät ja vertasin niitä kuhunkin kontrolloituun sanastoon. Suhteellisen osuuden laskin vertaamalla termien lukumäärää, jotka eivät ole kontrolloidussa sanastossa yksittäisten asiasanojen esiintymien kokonaismäärään.

15. Kentän 652 asiasanat

tietokanta	asiasanat, jotka eivät YSA:ssa	suhteellinen %-osuus
Lappi	13723	28,3
Kemi	3931	14,1
Keminmaa	1912	10,3
Posio	4727	27,6
Karpalo	18258	41,1

Musan ja Kaunokin asiasanojen vertailussa suhteellinen prosenttiosuus tarkoittaa termien, jotka eivät ole Musassa tai Kaunokissa eivätkä YSA:ssa, osuutta yksittäisten termien kokonaismäärästä.

16. Kentän 653 asiasanat

tietokanta	ei Musassa	löytyy YSA:sta	suhteellinen %-osuus
Lappi	4353	808	59,4
Kemi	610	339	6,3
Keminmaa	103	66	5,6
Posio	794	244	43,9
Karpalo	3921	509	69,8

17. Kentän 654 asiasanat

tietokanta	ei Kaunokissa	löytyy YSA:sta	suhteellinen %-osuus
Lappi	8553	1637	38,2
Kemi	7148	1177	39,8
Keminmaa	3406	667	21,5
Posio	7985	810	42,8
Karpalo	6474	1230	29,3

Yhteenveto

Tietokantojen analyysin tuloksena tuotin useita kymmeniä listoja erilaisista virheistä, joita tietokannoissa esiintyi. Se, mitä virheitä etsin ja miten, oli summa keskusteluista asiantuntijalähteiden (inhimillisten ja kirjallisten) kanssa yhdistettynä kirjastoista tulleisiin toiveisiin ja omaan pohdiskeluuni.

Erityisesti Usemarconin tuottamat virhelokit antavat kirjastolle hyvän kuvan siitä, mitä luettelointiformaatin konversiossa tapahtuu ja mihin seikkoihin tietokannan siivouksessa kannattaa kiinnittää huomiota. Testikonversio myös näyttää, että osa virheistä korjaantuu formaatin muutoksen yhteydessä.

Tietokantaa voi siivota ja virheitä korjata ennen konversiota. Tämä on toki mahdollista konversionkin yhteydessä. Tämä vaatii lisäyksiä ja muutoksia konversiosääntöihin. Esimerkkinä tästä analysoimissani tietokannoissa puuttuva 245-nimekekenttä. Suurimmassa osassa tapauksista nimeketieto on siirrettävissä konversion yhteydessä 240-kentästä.

Asiasanojen deduplikointi olisi tarpeellista tehdä, koska tulevassa konversiossa on tarkoitus yhdistää useampi tietokanta yhdeksi kokonaisuudeksi. Jos esimerkiksi eri kirjastokohtaisissa asiasanakentissä on sama termi, on yhdistymisen jälkeen yhdessä tietueessa monta samaa termiä.

Minitietueisiin ja tuplatietueisiin olisi myös hyvä kiinnittää huomiota ennen konversiota. Minitietueista monet voivat olla niin sanotusti roskaa syystä tai toisesta tietokantaan jääneitä puutteellisia tietueita. Tuplatietueiden etsintä vaatii aikaa ja kärsivällisyyttä, ainakin jos tuplien etsintää ei tehdä automaattisesti.

Monet listaamistani virheistä ovat sellaisia, että niistä on konversion jälkeen yhtä paljon tai vähän haittaa kuin tähänkin asti on ollut. Mutta esimerkiksi virheelliset vuosiluvut tai luokkamerkinnät voivat haitata ainakin tiedonhakua, joten niidenkin korjaaminen voisi olla paikallaan.

Lähteet

Asiantuntijalähteet

Matti Lassila, Osuuskunta Sange & Jyväskylän yliopiston kirjasto

Leena Kinnunen, Rovaniemen kaupunginkirjasto

Ilkka Leinonen, Kemin kaupunginkirjasto

Vesa Sarajärvi, Ranuan kunnankirjasto

Muut lähteet

FINMARC -> MARC 21 -konversiosuunnitelma bibliografisille tiedoille
http://www.kansalliskirjasto.fi/attachments/5m4XaGYjD/5sXYMv4U1/Files/CurrentFile/Finmarc---MARC21---bib---201009.pdf

Finmarc yhtenäisformaatti, 1998
http://www.kansalliskirjasto.fi/attachments/5m4XaGYjD/5yHU2H0dx/Files/CurrentFile/FINMARC98.pdf

Ojala, Mace 2012: *PIKI-tietokannan siivousmenetelmät, projektin loppuraportti*
http://www2.kirjastot.fi/File/6f93a390-3755-4795-ba22-2aef36b23cc6/pikisiivous.pdf

USEMARCON User Controlled Generic MARC Converter Manual
http://www.helsinki.fi/~ermaijal/usemarcon.pdf

Viitanen, Anna 2010: *Perussuunnitelma yleisille kirjastoille MARC 21 -formaattiin siirtymiseksi*
http://www.kirjastot.fi/sites/default/files/content/Perussuunnitelma.pdf