Vaskin luettelointiryhmä

25.10.2012

SUUNNITELMA KONVERSIOIDEN JÄLKIEN SIIVOAMISESTA VASKI-TIETOKANNASSA

**Miksi tietokantaa on siivottava**

Vaski-tietokannassa on viimeisten yhdistämiskonversioiden jäljiltä vanhojen ongelmien lisäksi uusia virheitä. Virheet heikentävät tietokannan käytettävyyttä ja vähentävät Vaskin kelpoisuutta kansalliseen metatietovarannon (Melindan) jäseneksi.

Sekä käytettävyyden että metatietovarantokelpoisuuden kannalta suurimman ongelman muodostavat tuplatietueet. Tietyillä kokoelman alueilla samasta nimekkeestä voi olla kahdeksan tietuetta. Koska näiden tietueiden kuvailutiedot ovat keskenään erilaisia, sekä asiakkaiden että henkilökunnan saamat tiedonhakutulokset ovat epäluotettavia.

Tietokannan käsityönä tehtävä siivous on välttämätön osa yhdistämiskonversioiden jälkihoitoa. Jos tietueiden yhdistäminen ei onnistu koneellisesti, myöskään siivoamista ei voida tehdä automaattisesti. Kevään konversion yhteydessä siivottiin tietokantaa niin paljon kun oli automaattisin menetelmin mahdollista. Niteettömät tietueet poistettiin tietokannasta ja lehtien tuplatietueista päästiin eroon. Muu siivous vaatii käsityötä. Erilaisia tietokannasta saatavia listauksia voidaan kuitenkin käyttää siivouksen apuna.

Vaski-tietokannassa ei tehty sen perustaneiden konversioiden jälkeen järjestelmällistä siivousta. Tietokannan tilaa alettiin kartoittaa ja tuplatietueita yhdistää vasta kun valmistautuminen viime kevään konversioihin alkoi. Muiden suurten tietokantojen esimerkki osoittaa, että konversioiden jälkeen siivouksen on oltava järjestelmällistä ja siihen on osoitettava resursseja.

PIKI-tietokannassa systemaattista siivousta on tehnyt 2000-luvulla yksi työntekijä useiden kuukausien ajan vuosittain. Oulun alueen Outi-kirjastoissa toteutettiin tietokantojen yhdistämiskonversio tämän vuoden alussa. Tietokannan siivous aloitettiin siellä heti konversion jälkeen ja sen pitäisi valmistua vuoden loppuun mennessä. Aikaa korjausten tekemiseen tulee Outissa tämänhetkisen arvion mukaan kulumaan noin 2800 tuntia. Outin tilanne poikkeaa Vaskin tilanteesta sikäli, että siellä ei konversion yhteydessä vaihdettu järjestelmää tai formaattia.

 Tässä suunnitelmassa esitellään Vaski-tietokannan siivoustarpeet, tarvittavat toimenpiteet ja arvio kunkin ongelman laajuudesta. Ehdotuksena on, että tietokannasta korjataan virheelliset aineistolajit ja tietueet joiden formaattikonversio on epäonnistunut. Paljon tuplatietueita sisältävät alueet käydään läpi ja korjataan järjestelmällisesti, lukuun ottamatta puheäänitteitä, elokuvia ja CD-ROM-levyjä. Tuplatietueiden korjaukset tehdään listojen perusteella. Työ ei yleisesti ottaen vaadi niteiden näkemistä ja sitä voidaan tehdä monessa Vaski-kirjastossa yhtä aikaa. Työtä kannattaa tehdä usean henkilön voimin. Korjausten tekijöillä on oltava luettelointiosaamista.

Korjauksiin tarvittavan työajan määrää on vaikea arvioida. Kokonaismäärä on luultavasti lähellä Outi-kirjastoissa korjauksiin käytettyä aikaa. Kunkin ongelman kohdalla on mahdollisuuksien mukaan esitetty arvio ongelman laajuudesta tietuemääränä tai arvio ongelman korjaamiseen kuluvasta ajasta.

Tietokannassa on nyt korjattavaksi esitettävien ongelmien lisäksi muitakin virheitä. Niitä on esitelty tämän suunnitelman lopussa

**Ensisijaisesti korjattavat virheet**

Virheet aineistolajeissa Aineistolajeja on korjattava Origo-kirjastoista peräisin olevissa tietueissa. Ongelmallisia ovat varsinkin äänitteet, koska niitä ei pystytty Origon aineistolajien perusteella konvertoimaan täysin oikein Auroraan. Aineistolajien korjaus ei Aurorassa ole helppoa, koska korjaajan on hallittava formaatin kiinteämittaiset koodit. Korjaukset kannattaa siksi tehdä keskitetysti. Samalla tarkistetaan ja korjataan virheellisten aineistolajien takia mahdollisesti syntyneet tuplatietueet.

Arvio korjausten vaatimasta ajasta: yhdeltä työntekijältä 2-4 päivää.

Tuplatietueet

* ISBN-numerolliset tietueet
	+ - Näitä tuplatietueita on jo siivottu ennen kevään konversiota poistamalla Turun monta ISBN-numeroa sisältäneitä tietueita ja yhdistämällä niistä johtuneita tuplatietueita.
		- Saman ISBN-numeron sisältävät tuplatietueet ovat erityisen tuhoisia seuraavia konversioita ajatellen.
		- Ongelman laajuudesta saadaan käsitys, jos tilataan Axiellilta listat tietueista, joissa olevat ISBN-tunnukset esiintyvät useassa tietueessa. Tällainen listaus sisältää myös turhaa tietoa, mutta näiden virheiden poistaminen ennen seuraavaa konversiota on välttämätöntä. Korjaukset tehdään listan perusteella.
		- Toinen vaihtoehto on tilata Axiellilta tietokantadumppi ja käyttää kirjastoista löytyvää koodausosaamista saman ISBN-numeron sisältävien tietueiden esille saamiseksi.
		- Tietokannassa on myös keskenään saman ISBN-numeron sisältäviä tietueita, joissa ei kyse tuplatietueista.
* ISBN-numerottomat kirjat
	+ - Ongelma koskee ennen 1970-luvun alkua ilmestyneiden kirjojen tietueita
		- Tapauksia on jo osittain käyty läpi ennen kevään konversiota (vuosina 1910-1933 ilmestyneiden kirjojen tietueet). Kevään konversiossa tuplatietueita tuli tällekin alueelle lisää.
		- Siivoamisessa ainoa vaihtoehto on tehdä hakuja julkaisuvuoden mukaan, järjestää tulos nimekkeen perusteella ja poimia joukosta tuplatietueet. Koska suurimmat ongelmat ovat uusimman aineiston joukossa, aloitetaan siivous vuodesta 1934.
		- Arvio ongelman laajuudesta: vuosien 1934 ja 1971 välillä ilmestyneiden kirjojen tietueita on tietokannassa 68 800. Tuplan sisältäviä nimekkeitä on 100 tietueen joukossa keskimäärin 10. Näissä tapauksissa tietueiden määrä / yksittäinen nimeke on 2-6.
* Musiikin CD-levyt
	+ - Tuplatietueita ei ole tähän mennessä käyty läpi.
		- Suurin tuplatietuemäärä on Turun ja Raision aineiston välillä. Järjestelmällisen läpikäynnin malli on samantapainen kuin ISBN-numerottomissa kirjoissa. Julkaisuvuotta ei kuitenkaan kannata käyttää hakutekijänä. Sen sijaan otetaan lähtökohdaksi Raision kokoelma.
		- Arvio ongelman laajuudesta: Raision nimekemäärä: 20 900
* Uudenkaupungin paikalliskokoelmat
	+ - Tietueet tuotiin Vaskiin tarkoituksella tuplatietueina, jotta niihin liittyvät asiasanat saatiin säilytettyä.
		- Uusikaupunki hoitaa näiden siivoamisen. Siivousta on jo aloitettu.
		- Ongelman laajuus: kokoelma sisältää 2000 nidettä, tuplatietueiden määrä on pienempi.
* Nuotit
	+ - Tuplatietueita ei ole tähän mennessä käyty läpi.
		- Nuoteissa käytetään vaihtelevasti ISBN- ja ISMN -tunnuksia, tietueiden yhdistäminen ei konversioissa siksi onnistu kovin hyvin standarditunnuksen perusteella.
		- Järjestelmällisen läpikäynnin malli on sama kuin ISBN-numerottomissa kirjoissa.
		- Arvio ongelman laajuudesta: nuotteja on tietokannassa 29 300, tuplatietueen sisältäviä nimekkeitä keskimäärin 6 /100 nimekettä. Näissä tapauksissa tietueen määrä / nimeke yleensä 2.
* AV-aineistot musiikin CD-levyjä lukuun ottamatta
	+ - Ennen kevään konversiota on yhdistetty äänikirjojen ja kielikurssien tuplatietueet. Ko. aineistolajeihin on tullut jonkin verran uusia tuplatietueita.
		- DVD, CD-ROM ja Blu-ray -levyjä, videoita ja moniviestimiä ei ole tähän mennessä käyty läpi.
		- Osa näiden aineistojen tuplatietueista tulee vastaan aineistolajien korjauksen yhteydessä. Lisäksi Kaarinassa on aloitettu kielikurssitietueiden siivoamista.
		- Koska kyseessä on melko nopeasti kuluva aineisto, näiden aineistolajien tuplatietueiden korjaaminen ei ole ensisijaista. Korjausten tekeminen on vaikeaa, koska luettelointitiedot ovat usein hyvin puutteellisia.

Tietueet, joiden formaattikonversio on epäonnistunut

76 tietueen tai osakohteen formaattikonversio on epäonnistunut. Nämä tapaukset löytyvät huomautuskenttään tallentuneen varoituksen tai virheilmoituksen perusteella. Suurin osa tietueista joudutaan luetteloimaan kokonaan uudelleen, osassa riittää pienempi korjaus.

Korjaukset on jo aloitettu. Ne tehdään Pallasta apuna käyttämällä vuoden loppuun mennessä Turussa.

Arvio korjausten vaatimasta ajasta: yhdeltä työntekijältä 2-3 päivää.

**Muita tietokannassa esiintyviä ongelmia**

Virheelliset kielikoodit Virheellisiä kielikoodeja on tietueissa paljon. Ne näkyvät verkkokirjaston laajennetun haun kielivalikossa. Korjaus on työlästä eikä tietokannan käytettävyyden ja tulevien konversioiden kannalta niin tärkeää kuin tuplatietueiden poisto. Pieni osa virheistä voidaan ehkä korjata massana. Tämä koskee lähinnä tikkukirjaimilla tallennettuja (muuten oikeassa muodossa olevia) koodeja.

Instrumentaalimusiikissa on käytetty kahta eri kielikoodia. Koodit voidaan yhdenmukaistaa massana. Auroran asetusten puutteellisten tiedonhakuominaisuuksien vuoksi korjausajo on tilattava Axiellilta.

Virheet asiasanoissa ja tekijätiedoissa

Asiasanoissa ja tekijätiedoissa on paljon vääriä muotoja. Väärät muodot aiheuttavat ongelmia tiedonhaussa, mutta myös mahdollisessa siirtymässä ontologiapohjaiseen järjestelmään. Näihin ongelmiin voidaan odottaa korjausapua Tampereella alkavasta tietokannan siivousmenetelmiä kehittävästä projektista. Projektissa saadaan toivottavasti rakennettua malleja, joiden avulla tietokannassa tietyissä kentissä olevia tietoja voidaan verrata ulkopuolisiin auktoriteettirekistereihin.