Jarmo Saarti, Kaisa Hypén, Arja Juntunen

Karkaavatko korpukset – muuttuva julkaisukulttuuri ja kirjallisuudentutkijan tiedonhallinta

Kirjallisuudentutkimuksen kohteena ovat tekstit ja niiden muodostamat kokonaisuudet. Tämä kuulostaa itsestäänselvyydeltä, mutta 1990-luvulta alkanut julkaisemisen digitalisoituminen asettaa tämän ajattelutavan koetukselle, erityisesti sen jälkimmäisen osan: tekstien muodostamat kokonaisuudet.

Samaan aikaan on tapahtunut julkaisumäärien raju kasvu: netcraft.com on arvioinut, että maailmassa on 663 miljoonaa Internet-sivustoa, ja GoogleBooks-projekti arvioi maailmassa julkaistun noin 130 miljoonaa kirjaa (Parr 2010). Tieteellisten artikkelien lukumäärä arvioidaan 50 miljoonaksi; lisäksi tulevat sellaiset aineistot kuten lehtiartikkelit ja arkistoaineistot. Internet on muodostanut myös uusia tekstityyppejä, jotka muuttuvat hyvin nopeasti. Tällaista määrää ei perinteisillä teknologioilla pystytä enää hallitsemaan.

Perinteinen, fyysinen tapa kohdata painetut tekstit ja niiden muodostamat korpukset on havainnollista. Kun näkee kirjaston tai arkiston hyllyt, on niiden kautta helppo hahmottaa kokonaisuus: metri aineistoa tai kymmenen hyllykilometriä kertoo selkeästi omaa viestiään tutkimuskohteesta. Digitaalinen aineisto ei aukene käyttäjälleen samalla tavoin, vaan sen hahmottaminen vaatii uudenlaisia teknologioita ja osaamista (vrt. myös Carrière & Eco 2011).

Suurimman haasteen digitaalisessa aineistossa asettaa dokumenttien ja niiden kontekstin ylläpitävä säilyttäminen. *Tulenkantajat*-lehdessä 1930-luvulla julkaistu runo saa lehdestä kontekstin sisällölleen siinä julkaistuista mainoksista, muista artikkeleista ja lehden taitosta. Miten vastaava konteksti näyttäytyy kahdeksankymmenen vuoden päästä verkkojulkaisuissa? Kuka tallentaa ja säilyttää samanlaiset kontekstit ja kuka säilyttää epävirallisen kontekstin, joka on Internetille ominaista? Jollei tätä tehdä, niin olemme luoneet julkaisukulttuurin, joka tekee kirjallisuudentutkimuksen ja erityisesti kirjallisuuden historian kirjoittamisen mahdottomaksi. Jos suomalaisesta kirjallisuudesta jää 2000-luvulta jäljelle vain painetut kirjat, antaa se vääristyneen kuvan kirjallisuudestamme sadan vuoden päästä katsottuna (ks. myös Ilva 2011).

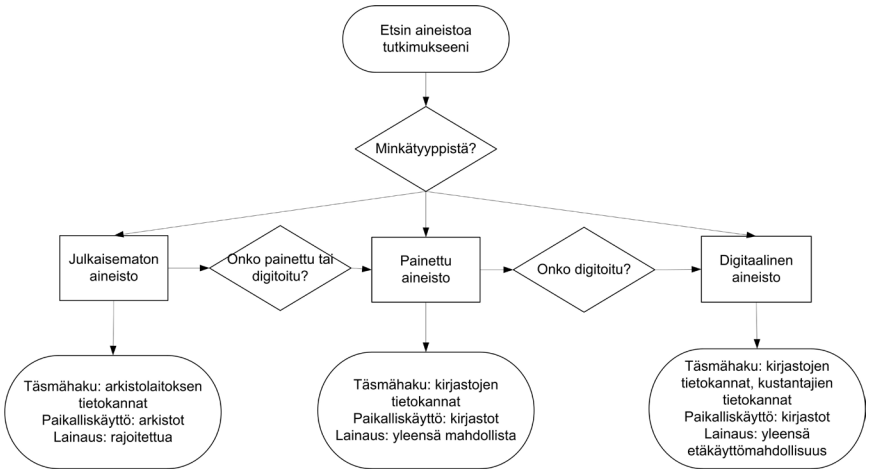
Tämän esseen tarkoitus on kahtalainen. Ensinnäkin esitellä alan tutkijoille keskeisimmät tiedonlähteet ja tiedonhaun työkalut, toiseksi pohtia kirjallisuudentutkimuksen tiedonhallintaa digitalisoituvassa kulttuurissa ja tarjota siihen työkaluja ja ideoita.

Painetut aineistot ja näiden digitointi

Dokumentteja on kerätty kokoelmiksi niin kauan kuin tekstejä on tuotettu. Kokoelmat jaetaan karkeasti kahtia: arkistoihin, jotka sisältävät ei-julkaistuja, uniikkeja tekstejä, sekä kirjastoihin, jotka koostuvat julkaistuista teksteistä. Käytännössä näiden kahden ero ei ole selkeä. Molemmilla organisaatioilla on omat tapansa tuottaa näistä dokumenteista järjestettyjä kokonaisuuksia, jotka mahdollistavat kokoelmien selailun sekä niihin kohdistuvan tiedonhaun (ks. kuva 1).

Suomalaiset kirjastojen painettujen aineistojen erikoiskokoelmat löytyvät ensinnäkin Kansalliskirjastosta, jolla on kattavin suomalaisen kirjallisuuden kokoelma. Sen lisäksi Kansalliskirjastossa on hyvä humanistisen alan tutkimuskirjallisuuden kokoelma. Vapaakappaleoikeuden omaavilla kirjastoilla (Jyväskylä, Oulu, Turku, UEF, Åbo) on myös laajat Fennica-kokoelmat.

Kansalliskirjastolla on laaja digitointiohjelma, jonka tavoitteena on digitoida kaikki suomalainen julkaistu aineisto. Tällä hetkellä digitoituihin aineistoihin laajimmat käyttöoikeudet on paikalliskäytössä vapaakappalekirjastoissa (mukaan lukien Kansalliskirjasto). Arkistolaitoksella on vastaavat rakenteet ja digitointiohjelmansa suomalaiselle arkistoaineistolle. Varastokirjastossa on myös laaja kokoelma vähän käytettyä painettua aineistoa.



Kuva 1. Aineistonhankinnan prosessi ja keskeiset lähdetyypit.

Tutkijan kannalta painettu aineisto ja sen fyysinen kohtaaminen auttavat hahmottamaan julkaisemisen kokonaisuuksia ja historiallisia muutoksia julkaisemisessa sekä suhteuttamaan oman tutkimuskohteen ja aineiston tähän kontekstiin. Digitoidut

aineistot antavat puolestaan uudenlaisia mahdollisuuksia humanistiselle tutkimukselle: yhdellä tiedonhaulla pystyy hallitsemaan laajoja korpuksia, joiden selailemiseen ei ihmisikä riittäisi.

Tutkijan tärkeimmät tietokannat

Tiedonhaun kannalta tietokannat voidaan jakaa karkeasti kahtia: viitetietokantoihin, joiden avulla etsitään painettua aineistoa, ja kokotekstitietokantoihin, jotka sisältävät dokumentit digitoituina. Viimeisen kymmenen vuoden aikana tietokantojen määrä ja tiedonhakuominaisuudet ovat lisääntyneet. Tutkijan oman työn tehostamisen takia kannattaa osallistua kirjastojen järjestämiin peruskursseihin tai tutustua huolellisesti tietokantojen omiin käyttöohjeisiin ja opetella viitetietojen tallentaminen.

Fennica, Suomen kansallisbibliografia, on luettelo suomalaisista julkaisusta (ks. www-linkit artikkelin lopussa). Siihen on tallennettu viitetiedot kirjoista vuodesta 1488 lähtien, lehdistä vuodesta 1771 lähtien ja lisäksi sarjajulkaisuja, karttoja, av- ja elektronista aineistoa. Fennicassa ovat myös ne ulkomailla julkaistut kirjat, joiden tekijä on suomalainen tai joiden aihe koskee Suomea.

Melinda-tietokannassa on Fennican lisäksi tiedot yliopistokirjastojen, Eduskunnan kirjaston, Varastokirjaston ja Tilastokirjaston aineistoista lukuun ottamatta e-aineistoa. Melinda on laajenemassa kattamaan ammattikorkeakoulujen ja yleisten kirjastojen kokoelmien viitetiedot.

Aleksi ja Arto ovat artikkeliviitetietokantoja. Aleksi-tietokannassa on viitteitä yleisaikakausilehdistä. Kirjallisuudentutkija voi käyttää sitä esimerkiksi kirjallisuusarvostelujen hakemiseen tai kirjailijahaastatteluiden etsimiseen. Arto-tietokantaan talletetaan kotimaisia tieteellisten ja ammattilehtien artikkeleita sekä kotimaisten kokoomateosten artikkeleita useilta tieteenaloilta.

Osa Arton artikkeleista on saatavilla kokotekstinä, jos lehti kuuluu Elektra-tietokantaan. Elektrassa on tällä hetkellä noin 30 kotimaista tieteellistä lehteä tai aikakauskirjaa. Usea yliopisto on perustanut vastaavankaltaisia julkaisuarkistoja, joiden kautta voi hakea ja lukea opinnäytetöitä, yliopiston tutkijoiden julkaisuja ja tutkimukseen liittyviä muita dokumentteja.

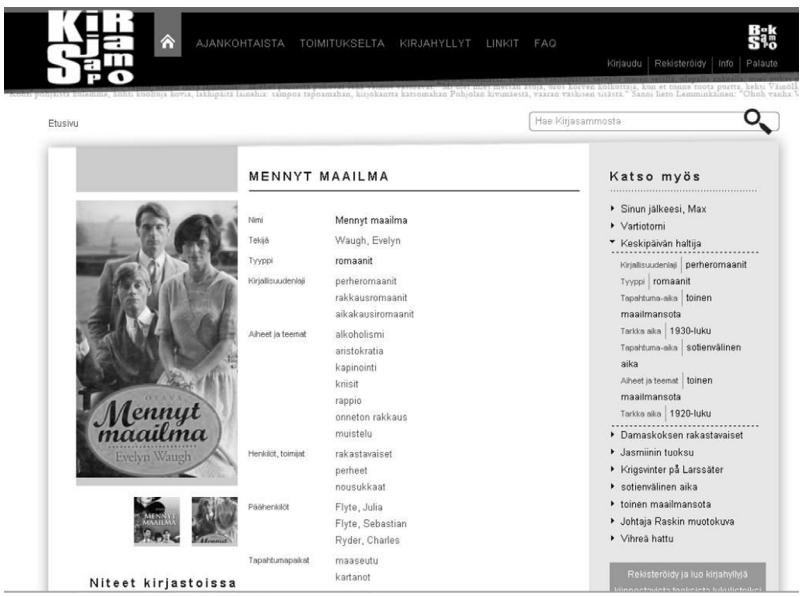
Nelli-portaali (National Electronic Library Interface) on kansallinen kirjastojen tiedonhakujärjestelmä, joka on käytössä suomalaisissa kirjastoissa ja jonka kautta voi hakea digitaalisia aineistoja. Yliopistojen Nelli-portaalit poikkeavat sisällöllisesti toisistaan sen mukaan, minkä tietokantojen tai elektronisten aineistojen lisenssit on kuhunkin yliopiston kirjastoon hankittu. Jos tietokantaan, e-lehteen tai e-kirjaan on käyttöoikeudet, ovat ne käytettävissä yliopiston henkilökunnalle ja opiskelijoille.

Yliopistojen kirjastot ovat rakentaneet Nelli-portaalinsa niin, että tietokannat ja lehdet löytyvät aihealueittain. Jos tietokantoja on useampia, on niistä tärkeimmät

nostettu listan alkuun. Haun voi tehdä samanaikaisesti useammasta tietokannasta, mutta viitemäärät voivat tällöin kasvaa suuriksi. Tällöin kannattaa siirtyä julkaisijan omaan käyttöliittymään ja hyödyntää sen hakua tarkentavia työkaluja. Parhaillaan ollaan luomassa kansallisen digitaalisen kirjaston palveluita, joiden tavoitteena on tehdä yksi käyttöliittymä, Finna, kaikkien suomalaisten arkistojen, museoiden ja kirjastojen aineistoihin. Todennäköisesti se korvaa edellä esitellyt erilliset tietokantojen käyttöliittymät. Hyvä perusohje on aloittaa tiedonhaku laajimmasta lähteestä – Suomessa Melindasta tai Finnasta.

Fiktioportaalit

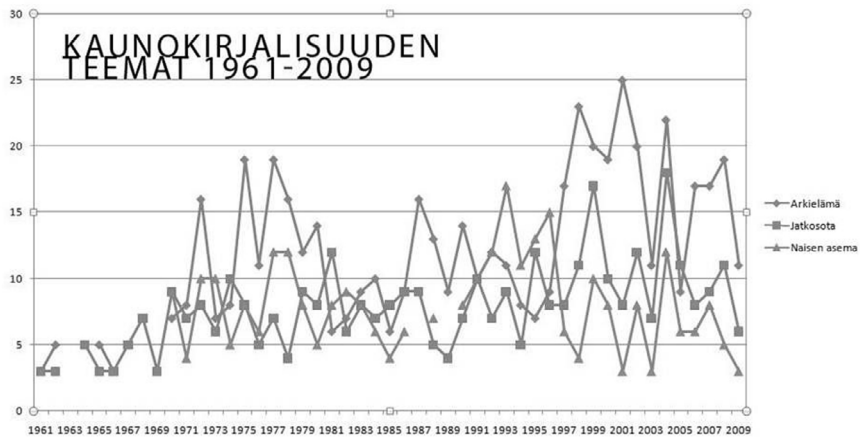
Internetissä on jo nyt useita eri toimijoiden laatimia kirjallisuusportaaleja, jotka sisältävät teosten analysointia laajemmin kuin perinteisissä kirjastojen tietokannoissa. Suomalainen Kirjasampo on yleisten kirjastojen ylläpitämä verkkopalvelu, jonne on tallennettu tietoja aikuisten suomen- ja ruotsinkielisestä kaunokirjallisuudesta. Kuvassa 2 on esimerkki Kirjasammon teoskuvailusta. Jokainen kuvailutermi on linkki, jolla saa esiin muita samalla sanalla kuvailtuja teoksia. Lisäksi järjestelmä esittää vastaavankaltaisia teoksia ja tarjoaa lisätietoja kirjailijasta hyperlinkkinä.



Kuva 2. Kirjasammon teoskuvailu ja teosten yhteyksien esittäminen.

Sisällönanalyysissä on pyritty välttämään liikaa tulkintaa. Tavoitteena on, että kuvailu kohdistuisi teosten helpoiten analysoitaviin elementteihin, kuten juoneen, teokses-

ta löydettäviin faktoihin ja selkeisiin teemoihin. Erityisen mielenkiintoiseksi kysymys kuvailun ”oikeellisuudesta” tulee, kun niiden pohjalta tehdään koneellisia analyysejä. Tämä on nähtävissä jo kuvan 2 mukaisessa esimerkissä. Vielä ilmeisemmäksi se tulee, kun tarkastellaan kuvan 3 grafiikkaa. Siinä on esitetty kolmen teeman – arkielämä, jatkosota, naisen asema – yleisyys kaunokirjallisuudessa vuosina 1961–2009. Taulukko perustuu Kirjasampoon tallennetun kuvailutiedon analysointiin. (Analyysivaiheessa ei 1960-luvun kirjallisuuden kuvailutyötä ole tehty systemaattisesti, siitä aukko.)



Kuva 3. Eräiden teemojen esiintyminen suomalaisissa kirjanimekkeissä 1961–2009 Kirjasammossa.

Kirjaston tuottama kuvailutieto on tässä astunut kokonaan uuteen rooliin. Se ei ole enää neutraali fakta, vaan sen avulla tehdään tulkintoja ja johtopäätöksiä kirjallisuuden kuvaamista ilmiöistä. Tämä herättää kysymyksen siitä, onko kirjastolaisten tuottama kuvailutieto riittävän luotettavaa? Voisiko esimerkiksi kuvassa kolme kuvatun kaltainen tilastollinen esitys olla osa kirjallisuudentutkimuksen taustamateriaalia? Tai voisivatko lukijakunta ja tutkijat tulla mukaan rikastamaan tallennettavaa dataa ja avaamaan uusia näkökulmia teoksiin?

Sisällönkuvailun välineet

Tiedontallennuksen keskeisiä työkaluja on asiasanoitus (tägitys, indeksointi). Asiasanoilla kuvaillaan teoksen keskeinen sisältö tiivistetysti tiedonhakua varten. Asiasanastoja on kehitetty systemaattisesti Suomessa erityisesti 1980-luvulta lähtien. Tutkija kohtaa asiasanoituksen hakiessaan itselleen aineistoa ja alkaessaan julkaista: useat kus-

tantajat vaativat tutkijalta itsenäistä käsikirjoitustensa asiasanoittamista.

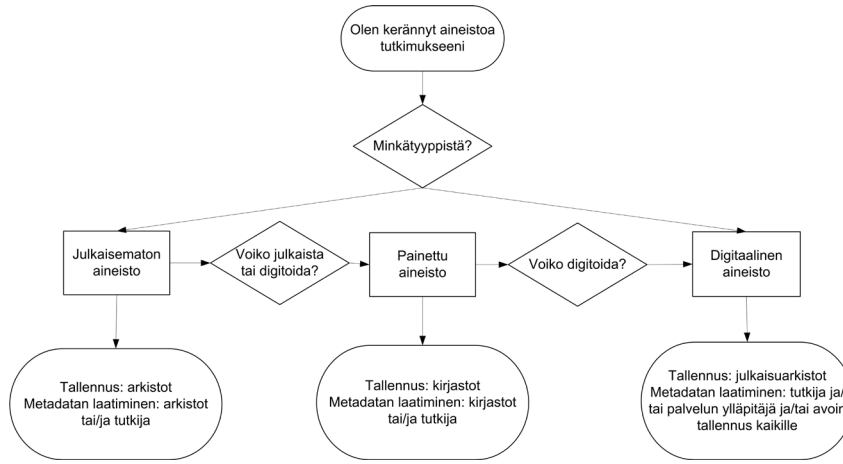
Tärkein suomalainen asiasanasto on Yleinen suomalainen asiasanasto (YSA) ja sen ruotsinkielinen versio ALLÄRS. YSA:n rinnalle on kehitetty eri alojen erikoissanastoja. Kirjallisuudentutkijan kannalta tärkeimmät erikoisalojen sanastot ovat fiktiivisen aineiston asiasanasto Kaunokki (Saarti & Hypén 2010) ja kirjallisuuden tutkimuksen asiasanasto Kitu. Näiden tietokoneiden tulkittavissa olevat, nk. ontologisoidut versiot ovat Yleinen suomalainen ontologia (YSO) ja KAUNO ja KITO (ontologioista tarkemmin ks. Hyvönen 2005, 2). Käyttämällä näitä välineitä tutkija varmistaa tuottamansa metadatan hyvän laadun ja parantaa aineistojen löytyvyyttä suomalaisista tietokannoista. Lisäksi suomalainen tiedeyhteisö on alkanut laatia omaa *Tieteen kansallista termipankkia*.

Tutkijan tiedonhallinta

Tiedonhallinta on keskeinen osa tutkimusprosessia. Suppeimmillaan se on oman tutkimuksen viitetietojen hallintaa ja hakemista, laajimmillaan se on huolen kantamista koko tieteenalan ja tieteen käytänteistä: se on osa hyvää tieteellistä käytäntöä, sen avulla hallitaan tutkimusaineiston elinkaari ja varmistetaan aineistojen jatkokäyttö ja avoin saatavuus (Yhteiskuntatieteellinen tietoarkisto 2012).

Tutkimuksen tiedonhallinnasta ja tutkimusdatan ja tutkimustulosten avoimesta julkaisemisesta on viime vuosina keskusteltu runsaasti. Myös Suomessa ja Euroopan Unionissa on otettu kanta, että kaikkeen julkisella rahalla tuotettuun tutkimukseen tulee liittyä tiedonhallinnan ja julkaisemisen suunnitelma ja että julkaisemisen tulisi olla avointa (ks. esim. Nuorteva 2008 ja Euroopan Komissio 2012). Tiedonhallinnasta on tullut siten osa jokaisen tutkijan työtä.

Nykyaikaiset teknologiat mahdollistavat tutkimuksen aineiston julkaisemisen ja jakelun usealla eri tavalla. Kuvassa 4 on esitelty oman aineiston julkaisemisen prosessi. Perinteisiä tapoja ovat aineiston tallentaminen johonkin arkistoon (esim. SKS) tai niiden julkaiseminen kirjana. Usealla yliopistolla on käytössä digitaaliset julkaisuarkistot, jotka mahdollistavat aineiston tallentamisen digitaalisena. Tällöin keskeiseksi tehtäväksi tulee metadatan tuottaminen aineiston löytyvyyden varmistamiseksi.



Kuva 4. Tutkimusaineiston säilyttämisen prosessi.

Kirjallisuudentutkija ei lue kirjoja vaan kirjallisuutta, ja tästä seuraa se, että alan tulee kantaa enemmän huolta tiedonhallinnasta ja siihen liittyvien välineiden kehittämisestä yhdessä muiden toimijoiden kanssa. Digitaaliset julkaisut muuttuvat ja katoavat nopeasti. Suomalaisen kirjallisuudentutkimuksen tulee pikaisesti laatia tutkimuskohteensa tallennussuunnitelma, muuten kohde jää hakijalta löytymättä tai saattaa jopa hävitä yllättävän nopeasti.

Lähteet

Digitaaliset tietokannat ja aineistot

Tietolähde:	**Keskeinen tietosisältö:**
Melinda (melinda.kansalliskirjasto.fi)	Suomalaisten kirjastojen aineistojen viitetiedot
Arto (arto.linneanet.fi) ja Aleksi (aleksi.btj.fi)	Suomalaisten lehtiartikkelien viitetiedot
Kansalliskirjaston digitoidut aineistot (digi.lib.helsinki.fi)	Suomalaisia aineistoja digitoituna
Finna (www.finna.fi)	Suomalaisten kirjastojen, museoiden ja arkistojen aineistot
Kirjasampo (www.kirjasampo.fi)	Suomalaisten kaunokirjojen ja niiden sisällön viitetietokanta
Kaunokki (kaunokki.kirjastot.fi/)	Fiktiivisen aineiston asiasanasto
Kitu (www.finlit.fi/kitu)	Kirjallisuuden tutkimuksen asiasanasto
Tieteen kansallinen termipankki (http://tieteentermipankki.fi/wiki/Termipankki:Etusivu)	Tiedeyhteisön laatima termitietokanta

Muut lähteet

CARRIÈRE, JEAN-CLAUDE & UMBERTO ECO 2011: *This Is Not the End of the Book. A Conversation Curated by Jean-Philippe de Tonnac*. Trans. Polly McLean. London: Harvill Secker.

EUROOPAN KOMISSIO 2012: Komission suositus tieteellisen tiedon saatavuudesta ja säilytettävyydestä. Annettu 17.7.2012. *Euroopan unionin virallinen lehti* L194, 39–43.

HYVÖNEN, EERO 2005: Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita? <http://www.seco.tkk.fi/publications/2005/hyvonen-miksi-asiasanastot-eivat-riita-2005.pdf > (12.1.2013)

ILVA, JYRKI 2011: Humanistit pilvessä: humanistitutkijat ja tieteellisen kirjaston rooli digitaalisessa maailmassa. *Tietolinja* (2). <http://www.kansalliskirjasto.fi/kirjastoala/tietolinja/0211/humanistitutkijat.html > (16.1.2013)

NUORTEVA, JUSSI 2008: Tiedonhallintasuunnitelma tehostaa tutkimusdatan käyttöä. *Tieteessä tapahtuu* 8/2008, 36–40.

PARR, BEN 2010: Google: There Are 129,864,880 Books in the Entire World. *Mashable Tech* 6.8.2010. <http://mashable.com/2010/08/06/number-of-books-in-the-world/ > (16.10.2012)

SAARTI, JARMO 1999: *Kaunokirjallisuuden sisällönkuvailun aspektit: kirjastoammattilaisten ja kirjastonkäyttäjien tekemien romaanien tiivistelmien ja asiasanoitusten yhdenmukaisuus*. Oulu: Oulun yliopisto.

SAARTI, JARMO & HYPÉN, KAISA 2010: From thesaurus to ontology: the development of the Kaunokki Finnish fiction thesaurus. *The Indexer* 2/2010, 50–58.

YHTEISKUNTATIETEELLINEN TIETOARKISTO 2012: *Tutkimusaineistojen tiedonhallinnan käsikirja* [verkkojulkaisu]. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. <http://www.fsd.uta.fi/tiedonhallinta/>. (15.10.2012)