Tutkimusdatan avaamisen esteet: haastattelututkimus Helsingin yliopistossa

Juuso Ala-Kyyny & Tuija Korhonen & Markku Roinila

Tiedepolitiikan ja rahoittajien linjauksissa on korostettu yhä enemmän tutkimusaineistojen avointa saatavuutta. EU:n Horisontti 2020 -ohjelma edellyttää aineiston avaamista ja Suomen Akatemia vaatii rahoitushakemuksissa aineistonhallintasuunnitelmaa. Myös monet julkaisijat vaativat yhä useammin tutkimusdatan liittämistä artikkeleihin. Avoimuuden vaatimus tieteessä ei toki ole uusi keksintö, tietotekniikan kehitys vain tarjoaa sille uusia käytännön sovelluksia. Mutta millaisia edellytyksiä tutkimusdatan avaamiseen ja jakamiseen käytännössä tällä hetkellä on? Tässä artikkelissa esiteltävä Helsingin yliopistossa tehty selvitys osoittaa, että nykyinen tutkimuskulttuuri ei ehkä kaikilta osin ole vielä valmis datan avaamiseen. Erityisenä ongelmana nousee esiin metadatan puutteellisuus.

Suomen yliopistot saivat keväällä 2017 opetus- ja kulttuuriministeriöltä (OKM) selvityspyynnön, joka koski tutkijoiden arvokasta tutkimuksen seurauksena syntyvää dataa ja sen siirtovalmiutta pitkäaikaissäilytykseen. Siirtovalmiudelle tarkoitamme tässä yhteydessä sitä, että tutkimuksessa syntynyt data on varustettu tarpeellisella metadatalla, jolloin sitä voidaan nimittää tutkimusaineistoksi. Selvityspyynnön taustalla on OKM:n vuosia valmistelema tutkimusaineistojen kansallinen pitkäaikaissäilytyspalvelu T-PAS (ks. Opetusministeriö 2008; Tutkimus-PAS-työryhmä 2017). Selvityksen tavoitteena oli saada kuvaa T-PAS-palveluun tarjottavien aineistojen määrästä ja laadusta.

Selvityspyyntö on kiinnostava myös Helsingin yliopiston (HY) näkökulmasta. Sen panostus tutkimusdatan hallintaan on kasvanut pari vuotta sitten hyväksytyn tutkimusdatapolitiikan myötä (ks. Helsingin yliopisto 2015). Tästä on osoituksena Mildred-projekti, jossa rakennetaan tutkimusdatan hallinnan infrastruktuuria HY:n tutkijoille (ks. Project Mildred 2017). OKM:lle tehtävä selvitys voidaan nähdä siis osana laajempaa tutkimusaineistoihin ja avoimeen tieteeseen liittyvää palvelukehitystä HY:ssä.

Tutkimuksen toteutus

Selvitystyön toteutti Helsingin yliopiston kirjaston tutkimuksen palveluiden työryhmä, johon kuuluivat tietoasiantuntija Mari Elisa Kuusniemi (työn ohjaaja) ja tietoasiantuntijat Juuso Ala-Kyyny, Tuija Korhonen ja Markku Roinila. Tutkimusmenetelmäksi valittiin haastattelut, koska kevättalvella 2017 toteutettu kyselytutkimus ei tuottanut tulosta.

Haastattelut toteutettiin kesällä 2017. Haastateltavina oli tiedekuntien dekaaneja ja tutkimuksesta vastaavia varadekaaneja sekä tutkijoita, tutkimuskoordinaattoreita ja tutkimuslaitosten edustajia. Eri tieteenaloilta oli haastateltavia suunnilleen samassa suhteessa kuin HY:ssä on tutkijoita eri aloilla. Kaikkiaan tehtiin 30 haastattelua neljällä HY:n kampuksella: keskustassa (11 haastattelua), Kumpulassa (6), Meilahdessa (7) ja Viikissä (5). Kampusjako on olennainen,

koska se ilmentää tieteenalaeroja, jotka vaikuttavat tutkimusaineistoihin: keskustassa ovat yhteiskunnalliset ja humanistiset tieteet sekä teologia, Kumpulassa matematiikka ja luonnontieteet, Meilahdessa lääketieteet ja Viikissä maatalous- ja metsätieteet, bio- ja ympäristötieteet, eläinlääketiede ja farmasia.

Keskitymme tässä artikkelissa haastatteluissa kartoitettuihin aineistonhallinnan yleisiin kysymyksiin: Miten eri tieteenaloilla tunnistetaan arvokas data? Miten sitä säilytetään ja miten se on kuvailtu? Millaisia mahdollisuuksia tai esteitä tutkimusdatan jatkokäytölle on? Kysymysten taustalla voi nähdä ajatuksen avoimesta datasta, ja olennaista onkin pohtia, miten hyvin nykyiset aineistonhallinnan käytännöt palvelevat tutkimusdatan jakamista, uudelleen- ja jatkokäyttöä. Haastatteluissa kysyttiin myös tutkijoiden näkemyksiä datan arvon määrittämisestä, mutta se jää laajuutensa vuoksi toiseen yhteyteen. Emme myöskään puutu suunnitellun pitkäaikaissäilytyksen käytäntöihin asian keskeneräisyyden vuoksi.

Käsitys tutkimusdatasta eri tieteenaloilla

Haastattelujen alussa muodostettiin kuva haastateltavan edustaman tieteenalan tutkimusdatasta. Jo tämä alustava kartoitus toi esiin tieteenalakohtaisia eroja haastateltavien välillä, ja erot korostuivat, kun tarkasteltiin säilyttämisen arvoisia aineistoja. Tulos ei sinänsä yllättänyt: tieteenaloilla, joilla tutkimustyö tuottaa paljon dataa, käsitys oman alan tutkimusdatasta oli parempi kuin aloilla, joissa ei tiedosteta, että tutkimus tuottaa tutkimusjulkaisun lisäksi myös dataa.

Keskustakampuksen tieteenaloilla datan säilyttämistarve liittyi usein historiallisen muistin vaalimiseen tai yhteiskunnalliseen merkitykseen. Muilla kampuksilla taloudelliset sekä terveyteen (Meilahti) liittyvät arvot tulivat voimakkaasti edellä mainittujen rinnalle. Yleisesti ottaen säilyttämisen arvoiseksi mainittiin laajat pitkän aikavälin kattavat analogiset ja digitaaliset aineistot sekä kansainväliset ja kalliilla rahalla toteutettujen tutkimusten aineistot.

Meilahden, Kumpulan ja Viikin kampuksilla tutkimuksesta syntyi lähinnä erityyppistä raakadataa tai siitä johdettua dataa (mm. analyysidata). Humanistis-yhteiskuntatieteellisesti painottuneen keskustakampuksen ero kovia tieteitä edustaviin kampuksiin oli selvä. Keskustan haastatteluissa säilyttämisen arvoisia aineistoja pohdittiin usein vasta haastattelutilanteessa – osa haastateltavista koki, että näillä tieteenaloilla ei edes tuotettaisi tutkimusdataa. Keskustakampuksen sisällä oli toki myös huomattavia eroja. Esimerkiksi suomen kielen oppiaineessa oli tarkka kuva säilyttämisen arvoisista kieliaineistoista.

Käytettävissä olevat säilytysratkaisut vaikuttivat käsitykseen tutkimusaineistoista. Suomen kielen oppiaineen kieli- ja murreaineistot tarjoavat jälleen hyvän esimerkin: ne tunnettiin keskimääräistä paremmin, koska niille on säilytyspaikka FIN-CLARIN-konsortion ylläpitämässä Kielipankissa.

Säilytysratkaisujen ohella kansainvälinen tutkimusyhteistyö oli yleensä merkki siitä, että myös käsitys alan tutkimusaineistoista oli keskimääräistä paremmalla tasolla. Esimerkiksi Kumpulan kampuksella on useita kansainvälisesti merkittäviä tutkimusaineistoja, joita säilytettiin osin Suomen ulkopuolella. Biogeotieteissä on nisäkkäitä koskeva havaintoaineisto (NOW-tietokanta), tähtitieteessä Planck- ja Euclid- tutkimushankkeiden aineisto (Euroopan avaruusneuvosto ESA) ja fysiikan CMS-kokeessa syntyvä aineisto (Euroopan hiukkasfysiikan tutkimuskeskus CERN).

Puutteellinen metadata pitkäaikaissäilytyksen kompastuskivenä

Kaikilla tieteenaloilla tutkimusdatasta vastaavat pääsääntöisesti tutkijat tai tutkimusryhmät itse. Keskustakampuksella joillakin tieteenaloilla on tehty sopimuksia tutkimusdatan käsittelystä, mutta useimmiten sen luovuttamisesta vas-

taa alalla vallitsevan käytännön mukaan vastuullinen tutkija.

Digitaalisia aineistoja säilytetään vaihtelevan huolellisesti, mm. koneiden kovalevyillä, henkilökohtaisilla verkkolevyillä, yliopiston yhteiskäytössä olevilla verkkolevyillä, pilvipalvelimilla, muistitikuilla, ulkoisilla kovalevyillä ja biopankeissa (Kumpulan ja Meilahden kampus). Joillekin tutkimusaineistoille on olemassa joko sovittu pitkäaikaissäilytysratkaisu tai suunniteltu ratkaisu, mikäli ao. instituutiot (kuten Kansallisarkisto tai Tietoarkisto) ottavat aineiston vastaan.

Suurin osa tutkimusdatasta sijaitsee Kumpulan, Meilahden ja Viikin kampuksilla, mutta myös keskustakampukselta sitä löytyy jonkin verran. Keskustakampuksella ja Meilahden kampuksella suuri määrä aineistosta on analogisessa muodossa, mutta sitä löytyy jonkin verran myös Viikistä ja hieman Kumpulasta. Tällaisia ovat mm. biopankkinäytteet, kasvinäytteet, litteroidut haastattelut, esineet ja VHS-nauhat. Useimmat haastateltavat katsoivat, että analogisen aineiston voi hävittää, jos aineisto digitoidaan. Samalla tuli kuitenkin ilmi, että hävittämiskäytäntöjä on vielä kehitettävä siten, että tutkimusdatan sensitiivisyys otetaan huomioon.

Digitaalinen data on useimmilla keskustakampuksen tieteenaloilla yleisesti käytössä olevissa tiedostomuodoissa, mutta Meilahdessa ja Viikissä on jonkin verran laiteriippuvaista dataa ja kaikilta kampuksilta löytyy jonkin verran ei-standardeja tiedostoformaatteja.

Metadata oli haasteellinen asiakokonaisuus monille haastatelluille. Siitä oltiin eniten tietoisia aloilla, joilla on muodostettu selkeät säilyttämisratkaisut. Monet kyllä tunnistivat sisällönkuvailun tärkeyden ja paikoin toivottiin koulutusta asian suhteen, jotta tutkijat voivat liittää metadatan tutkimusmateriaaliin tutkimuksen kuluessa. Useimmiten metadatan muodostaminen nähtiin kuitenkin joko ylimääräisenä työnä johon ei ole aikaa tai tehtävänä, joka on jonkun muun kuin tutkijan itsensä, kuten tutkimusavustajan, vastuulla.

Haastatteluissa tuli selvästi esiin, että metadata onkin enimmäkseen hyvin puutteellista eikä sen tuottamiseen tai parantamiseen ole resursseja. Lisäksi pitkäaikaissäilytykseen luovutusta varten tarvitaan resursseja analogisen aineiston digitoimiseen ja digitaalisen datan siivoamiseen. Osa näistä digitoimistarpeista on kiireellisiä, sillä erityisesti Viikissä oltiin huolestuneita aineiston katoamisesta tutkijoiden eläköitymisen yhteydessä.

Dokumentoinnin puutteellisuus nousee todennäköisesti kynnyskysymykseksi, kun aineistoa valitaan T-PAS-palveluun. Aineistot joissa metadata on kunnossa, ovat jo pääosin muualla pitkäaikaissäilytyksessä. Haastatteluissa tuli myös esiin useita tapauksia, joissa tutkija katsoi aineiston olevan siirrettävissä, mutta myöhemmin selvisi, että sen sisällönkuvailu vaatii vielä töitä.

Jatkokäytön haasteita

Tutkimusdatan sensitiivisyys korostui Meilahden kampuksen aineistojen kohdalla. Lääke- ja terveystieteessä ollaan hyvin tietoisia siitä, että arkaluontoisia henkilötietoja on käsiteltävä eettisesti jo lainsäädännöllistäkin syistä. Niitä säilytetäänkin lukittujen ovien takana tai suljetussa verkossa salasanojen takana. Sensitiivisyyden takia tutkimusaineiston anonymisoinnissa on haasteita. Esimerkiksi Meilahdessa on aineistoa, josta henkilön voi tunnistaa sadan prosentin varmuudella ja sen käsittelyyn on mietitty ratkaisuja mm. tulevan Genomikeskuksen yhteydessä (ks. STM tiedote 49/2016).

Keskustakampuksella sensitiivistä dataa on lähinnä tutkijan muistelmissa ja haastatteluaineistoissa. Joillakin aloilla anonymisointi voi olla nopeaakin, mikäli datan rakenne on selkeä ja nimet esiintyvät vain tietyissä kohdisssa. Kontekstin poistaminen voi toisaalta vähentää aineiston tieteellistä arvoa: jos lupa säilyttämiseen on pyydettävä jokaiselta aikoinaan haastatellulta, aineiston kattavuus voi kärsiä.

Kumpulassa ja Viikissä ei juuri ole sensitiivistä dataa, mutta usein tutkijat haluavat pitää datan vain omassa käytössään siihen asti, kun tutki-

mus on tehty ja julkaisu ilmestynyt. Joillakin tutkimusaloilla, kuten metsäntutkimuksessa ja lääketieteessä, on patentteihin liittyviä suoja-aikoja tutkimusdatan avaamiseksi.

Monilla tieteenaloilla tutkimusdata haluttiin pitää käytettävissä, vaikka se siirrettäisiin pitkäaikaissäilytykseen, sillä se on tutkijalle edelleen relevanttia tutkimusmateriaalia. Kumpulan, Meilahden ja Viikin kampuksilla tutkijat olivat hyvin tietoisia siitä, että tutkimusdataa vaaditaan yhä enenevissä määrin avoimiksi, esimerkiksi julkaisujen yhteydessä. Tähän suhtauduttiin periaatteessa myönteisesti ja nähtiin, että vanhasta aineistosta voi tulevaisuudessa saada irti uutta tietoa. Humanistis-yhteiskuntatieteellisillä aloilla nähtiin julkaisut avoimen tieteen keskeisimmäksi muodoksi. Haastatteluissa kävi myös ilmi, että tutkimusaineiston pitkäaikaissäilytys ei ollut kaikille tutkijoille selvä käsite ja sen määrittelyä pyydettiin usein selittämään.

Loppupäätelmät

Haastattelututkimus osoitti, että tutkimusdatan avaaminen ja jatkokäytön huomioiminen, omaa tai oman tutkimusryhmän tutkimustyötä laajemmassa mielessä, edellyttää nykytilanteessa useimmiten ulkoista painetta. Aineistonhallintaa edesauttaa, jos säilytysratkaisu on olemassa – ja jos säilytykseen siirtoa varten on olemassa vakiintuneet toimintatavat.

Kun julkisin varoin tuotettu tutkimusdata halutaan avata muiden tutkijoiden ja aktiivisten kansalaisten käyttöön, sen kuvailuun ja metadataan on kiinnitettävä paljon nykyistä enemmän huomiota. Siihen tarvitaan koulutusta, välineitä ja sopivia kannustimia. Tämä oli selvityksemme tärkein havainto. Jo etukäteen oli tiedossa, että aineistonhallinnassa ja metadatassa on puutteita, mutta puutteiden laajuus tieteenalasta riippumatta yllätti selvityksen tekijät. Tutkimusaineiston elinkaariajattelu ei selvityksen valossa näytä kovin keskeiseltä osalta tutkimuskulttuuria.

Selvityksen tulokset ovat samansuuntaisia kuin kymmenen vuoden takaisessa Sami Borgin ja Arja Kuulan (2007, 70) raportissa, joka koski OECD:n datasuositusten toimeenpanomahdollisuuksia Suomessa: ”Yli puolet professoreista koki sähköisten aineistojen jatkokäytön tärkeäksi esteeksi sen, että aineistojen tietosisällöt ja tiedostot ovat puutteellisesti dokumentoituja. Noin joka toinen arvioi aineistojen avoimuudessa haitaksi sen, että tutkijoiden työaikaa menisi vanhojen aineistojen muokkaamiseen käyttökelpoisiksi.”

Myös HY:n selvityksessä tutkijat kokivat dokumentoinnin resurssikysymyksenä, ja tutkimusaineiston kuvailu nähtiin ylimääräisenä työnä, joka tulee varsinaisen tutkimus- ja julkaisemistyön päälle. Moni tutkija katsoi, että tämä voidaan jättää esimerkiksi tutkimusavustajan tai opiskelijan vastuulle. On kuitenkin huomattava, että dokumentointi kuuluu olennaisesti tutkimusdatan hallintaan ja sen voi tehdä luotettavasti vain tutkija itse ja nimenomaan tutkimusprosessin aikana – jälkikäteen työ on paljon vaikeampaa ja tulokset ovat huonompia, sillä konteksti on jo hämärtynyt ja yksityiskohdat ovat osin unohtuneet. Tämän oli oivaltanut vain muutama tutkija.

Mielestämme selvitys osoittaa, että tutkimusdatan avaamisen esteiden raivaamiseksi tutkijoiden olisi omaksuttava uudenlainen toimintakulttuuri, jossa tutkimusdatan dokumentointi aloitetaan heti tutkimusprosessin alkaessa. Kirjastot voivat tarjota metadatan muodostamiseen apua ja koulutusta, mutta tutkimusdatan dokumentointi on nähtävä kiinteänä osana tutkimusprosessia, joka kuuluu tutkijan vastuulle.

Lähteet:

Borg, Sami & Arja Kuula (2007). Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari. Valmisteluraportti OECD:n datasuosituksen toimeenpanomahdollisuuksista Suomessa. Tampereen yliopisto. Yhteiskuntatieteellisen tietoarkiston julkaisuja; 6. Saatavana verkossa: <http://www.fsd.uta.fi/fi/julkaisut/julkaisusarja/FSDjs06.html> [viitattu 20.11.2017].

Helsingin yliopisto (2015). Tutkimusdatapolitiikka. Helsingin yliopiston tutkimusdatapolitiikka, hyväksytty 11.2.2015. Saatavana verkossa: < http://www.helsinki.fi/kirjasto/fi/avuksi/tutkimusdatan-hallinta/tutkimusdatapolitiikka/> [viitattu 17.11.2017].

Opetusministeriö (2008). Sähköisen aineiston pitkäaikaissäilytystä ja käyttöä koskevan työryhmän muistio. Opetusministeriön työryhmämuistioita ja selvityksiä 2008:2. Helsinki: Opetusministeriö, Koulutus- ja tiedepolitiikan osasto. Saatavana verkossa: <https://julkaisut.valtioneuvosto.fi/handle/10024/79392> [viitattu 17.11.2017].

Project Mildred (2017). Mildred-projektin blogisivusto. Saatavana verkossa: <http://blogs.helsinki.fi/mildred/> [viitattu 20.11.2017].

Sosiaali- ja terveysministeriön tiedote 49/2016; http://stm.fi/artikkeli/-/asset_publisher/genomikeskus-tuoperimasta-saatavan-tiedon-osaksi-terveydenhuoltoa.

Tutkimus-PAS-työryhmä (2017). Tutkimusaineistojen tiedostomuodot ja pitkäaikaissäilytyskelpoisuus. Avoin tiede ja tutkimus -hankkeen selvityksen loppuraportti 10.2.2017. Saatavana verkosta: < https://avointiede.fi/documents/10864/12232/Tutkimusaineistojen_tiedostomuodot_loppuraportti.pdf/24557e81-f504-4383-9a27-304e09b27e94> [viitattu 23.11.2017].

Tietoa kirjoittajista:

Juuso Ala-Kyyny, tietoasiantuntija
Helsingin yliopiston kirjasto
juuso.ala-kyyny@helsinki.fi

Tuija Korhonen, tietoasiantuntija
Helsingin yliopiston kirjasto
tuija.korhonen@helsinki.fi

Markku Roinila, tietoasiantuntija
Helsingin yliopiston kirjasto
markku.roinila@helsinki.fi