KATSAUS

Tutkimusdata tieteellisenä julkaisuna

Mari Elisa Kuusniemi
Helsingin yliopiston kirjasto
mari-elisa.kuusniemi@helsinki.fi
https://orcid.org/0000-0002-7675-287X

Susanna Nykyri
Tampereen yliopiston kirjasto
susanna.nykyri@tuni.fi
https://orcid.org/0000-0002-5018-5176

Molemmat kirjoittajat osallistuneet tasapuolisesti.

In this article we drill down to the concept of data publication. We introduce how the term research data publication is defined in academic settings and literature. We discuss how research data is published, how the data publications are reviewed or curated, and which are the current incentives for the publication of research data. These are the main questions we aim to answer to in this article.

As conclusion we ponder how we can learn from the best practices already in place in academic communities and how to spread these further to areas which have not yet created their communication culture to cover data publications. Finally, we illustrate how we could support the development as information and data management experts.

Asiasanat: avoin tieto, yhdistetty avoin tieto, metadata, tieteellinen julkaisutoiminta, tutkimusaineisto, kannustimet

Pysyvä osoite: https://doi.org/10.23978/inf.109094

Johdanto

Avoin tiede on avointa julkaisemista paljon laajempi käsite. Tutkimusaineistojen hyvä hallinta ja vastuullinen avaaminen on olennainen osa avointa tiedettä. Avoimuuden suurimpina hyötyinä pidetään alkuperäisten tietolähteiden saatavuutta, tutkimuksen laadun varmistamista ja moninaisten näkökulmien tarjoamista tietoon. Tutkimustuotosten ja -menetelmien avoimuus on myös yhä enemmän tutkijaa meritoivaa toimintaa. EU:n lainsäädäntö säätelee jo julkisin varoin tuotettujen data-aineistojen avoimuutta ja kansallinen lainsäädäntö on luonnollisesti seuraamassa tätä kehitystä. Nähtäväksi jää, miten tämä tulee konkreettisesti vaikuttamaan tutkimusdatan julkaisemiseen. Tutkimuksen digitalisaatio tuo uusia mahdollisuuksia tutkimustuotosten avoimuudelle ja tämän kehityksen myötä tutkimusta ohjaavat toimintaperiaatteet, lainsäädäntö (ks. esim. Hallituksen esitys eduskunnalle avoimen datan direktiivin täytäntöönpanoa koskevaksi lainsäädännöksi, 2020) ja mittarit muuttuvat.

Pelkkä avoin julkaiseminen ei vielä yksinään riitä. Esimerkiksi löydettävyyden ja saatavuuden edistämiseksi on tehtävä paljon muutakin kuin vain julkaistava pdf-muotoinen teksti tai datasetti internetissä. Tässä kirjoituksessa tarkastelemme syitä julkaista dataa, pohdimme tutkimusdatan käsitettä ja mitä on tutkimusdatan vertaisarviointi, ja lopuksi esitämme johtopäätöksiä mitä käsittelemämme tutkimusdatan julkaisunäkökulma merkitsee tutkijoiden palveluiden kehittämiselle erityisesti yliopistokirjastoissa.

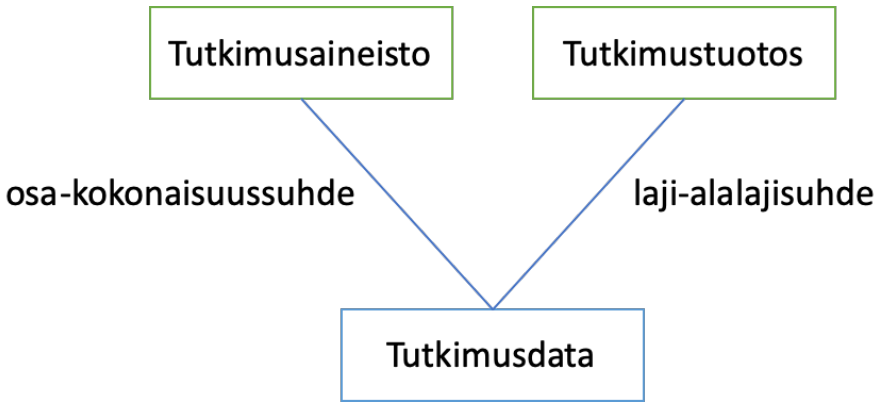
Datasta tutkimusdataan

Mitä on tutkimusdata (eng. *research data*)? Onko se eri asia kuin tutkimusaineisto (eng. *research material*)? Tähän ei ole vakiintunutta määritelmää, ja vaihtoehtoisia tarkastelukulmia on useita (ks. esim. YSO 2021, Tieteen termipankki 2021, Parland-von Essen & al. 2018). Näitä sanoja käytetään ristiin toistensa synonyymeina tai ne määritellään ja käännetään kontekstista riippuen eri tavoin.

Tässä artikkelissa tarkoitamme tutkimusdatalla sitä tietoaineistoa, joka on kerätty, havaittu, mitattu tai luotu tutkimustulosten todentamiseksi. Konteksti tekee datasta tutkimusdataa. Konteksti on tässä tapauksessa tutkimustarkoitus. Mikä tahansa tietoaineisto (data) voi olla tutkimusdataa, jos sitä analysoidaan tutkimustarkoituksessa.

Tutkimusaineisto on tutkimusdataa laajempi käsite. Se pitää sisällään tutkimusdatan lisäksi tutkimuksen sekundäärilähteet, kuten käytetyn kirjallisuuden ja menetelmäohjeet. Myös fyysiset artefaktit ja näytteet ovat osa tutkimusaineistoa.

Samoin tutkimustuotos on tutkimusdataa laajempi käsite. Tutkimusdata on yksi tutkimustuotostyyppi. Näin tutkimusdata on geneerisessä suhteessa (laji-alalaji) tutkimustuotokseen nähden ja partitiivisessa suhteessa (osa-kokonaisuus) tutkimusaineistoon nähden. (Suhteista, ks. ISO 25964 – *the International standard for thesauri and interoperability with other vocabularies*.)



Kuva 1: Tutkimusdatan suhde käsitteisiin tutkimusaineisto ja tutkimustuotos

Tutkimusdataa tuotetaan, kerätään ja käytetään ainakin empiirisessä tutkimuksessa. Teoreettisen tutkimuksen datan tunnistaminen on hankalampaa ja saattaa olla, että sitä ei aina synny jaettavassa tai julkaistavassa muodossa. Käytännössä tutkimus on usein monimenetelmäistä ja tutkijoidenkin on välillä vaikea hahmottaa kaikkea sitä tutkimusdataa mitä he käyttävät ja tuottavat. Tämä johtuu osittain tutkimusperinteestä, jossa tutkimuksen tuotoksena on tyypillisesti nähty artikkeli, kirja, konferenssijulkaisu tai muu proosa. Näiden perinteisten tutkimusjulkaisujen ympärille ovat syntyneet tarvittavat kannusteet, prosessit ja palvelut. Nyt kun tutkimusdataa voidaan digitalisaation myötä helpommin jakaa ja julkaista sekä uudelleen- ja jatkokäyttää, myös se aletaan enenevästi nähdä tutkimustuotoksena. Samalla se myös tunnistetaan paremmin. Joillakin aloilla tutkimusdatan jakaminen on kiinteä osa tieteellistä julkaisemiskulttuuria jo nyt, mutta monilla muutos tähän suuntaan on vasta aluillaan.

Arvokas ja laadukas tutkimusdatatuotos ei synny vahingossa tutkimusprosessin sivutuotteena (ks. esim. Laine 2018, Laine & Nykyri 2018). Kuten artikkelin kirjoittaminen vaatii taitoa esittää asia selkeästi ja ryhmitellä sanottava ymmärrettävään rakenteeseen, myös tutkimusdata on arvokasta vain, jos se on selkeässä ja ymmärrettävässä muodossa. Tutkimusdataa ei useinkaan voi järjestää ja korjailla kuntoon tutkimusprojektin jälkeen, vaan sen rakenne syntyy osana tutkimuksen suunnittelua ja toteutusta. (Tämä artikkeli ei käsittele tarkemmin datan rakenteen syntyä osana tutkimusprosessia. Siitä asiasta kannattaa lukea informaatiotutkimuksen kielellä esimerkkejä Lisa M. Givenin ja Hope A. Olsonin artikkelista (2003) *Knowledge organization in research: A conceptual model for organizing data*.) Parhaassa tapauksessa tutkimusdata ei ole ymmärrettävää vain ihmislukijalle, vaan se on ymmärrettävää myös koneelle (*FAIR principles*, Wilkinson & al. 2016). Silloinkin kun tutkimusdataa syntyy tutkimusprosessissa melko automaattisesti, niin laadukkaaksi tutkimustuotokseksi tutkimusdata ei muutu itsestään. Datatuotos vaatii tarkkaa suunnittelua, tiedon rakenteen hahmottamista, standardien noudattamista ja ahkeraa dokumentointityötä. Hyvin suunniteltu tutkimusprosessi tuottaa datatuotoksia, joita voidaan kutsua myös data-aineistoiksi (eng. *data set*).

Mitä on tutkimusdatan julkaiseminen?

Tutkimusdata on perinteisesti julkaistu osana tutkimusjulkaisua; artikkeleissa ja kirjoissa joko tekstin lomassa tai liitteinä. Yhä useammat tiedekustantajat edellyttävät tieteellisten artikkeleiden taustalla olevan tutkimusdatan täydellistä tai rajattua avaamista ehtona julkaisulle. Datan julkaiseminen osana tieteellistä julkaisua lieneekin nykyään yleisin tutkimusdatan julkaisemisen tapa. Varsinaisia datajulkaisuja taas julkaistaan tähän erikoistuneissa lehdissä eli ns. data journaleissa (*data journal*) tai tutkimusdatan avaamiseen suunnitelluissa datatietokannoissa, joita kutsutaan datarepositorioiksi tai -arkistoiksi. Näissä data voidaan julkaista joko kiinteästi liitettynä perinteiseen tiedejulkaisuun tai itsenäisenä kokonaisuutena. Myös internetiin, esimerkiksi projektin verkkosivuille saataville asetettu tutkimusdata tai tietokanta voi olla datajulkaisu. Esimerkiksi tutkimusinfrastruktuurien mittalaitteiden tuottamia aikasarjoja jaetaan verkossa ja ne ovat kansainvälisesti tärkeitä tiedonlähteitä. Tutkimusdatan erilaisia julkaisutyyppejä ovat kuvanneet artikkelissaan Lawrence & al. (2011).

Klump & al. pohtivat datajulkaisun käsitettä artikkelissaan *Data publication in the open access initiave* (Klump & al. 2006). He asettavat datajulkaisulle seuraavat kriteerit:

A. Datajulkaisulla tulee olla pysyvä tunniste, jonka avulla dataan voidaan viitata.
B. Datan tulee olla käyttökelpoista ja laadukasta.

Avoimen tieteen periaatteiden mukaisesti dataan pitää myös olla pääsy ja sen lisenssin on mahdollistettava datan käyttö. (emt.) Tuoreet vaatimukset (esim. Suomen Akatemia) ja linjaukset (esim. Tutkimusaineistojen ja -menetelmien avoimuus. Korkeakoulu- ja tutkimusyhteisön kansallinen linjaus ja toimenpideohjelma 2021–2025. Osalinjaus 1: Tutkimusdatan avoin saatavuus) tähdentävät datan keskeisen metadatan avoimuutta silloinkin, kun itse tutkimusdata ei ole täysin tai osin avattavissa.

Viitattavuus on datajulkaisun tärkeä ominaisuus (Laine & al. 2018). Dataan viittaamisen edellytys on jonkinlainen suhteellisen pysyvä tunniste, jonka avulla pääsee datan metatietoihin. Tunniste voi olla minimissään verkkosivun osoite eli URL, mutta pysyvän tunnisteen (DOI, URN, handle) avulla viittaus on pidempään jäljitettävissä (e-viittaamisesta, ks. Hakala 2017 ja tunnisteen valinnasta Lahtinen & al. 2020). Dataan viittaaminen ei sinällään eroa muuhun tutkimusaineistoon viittaamisesta, mutta käytännöt kuitenkin vaihtelevat melko tavalla. Käytäntöjen yhtenäistämiseksi tiedekustantajat ovat alkaneet antaa ohjeita dataan viittaamisesta. Tiedekustantajia ohjataan myös laatimaan datapolitiikka tukemaan julkaisun taustalla olevan datan asianmukaista hallintaa silloinkin, kun julkaisun pääasiallinen sisältö on muu kuin data, mutta tutkimustulokset perustuvat ainakin osittain tutkimusdataan (ks. Ilva & al. 2020).

Ollakseen käyttökelpoista, datan täytyy olla hyvin dokumentoitua. Hyvä dokumentointi tuottaa rikasta metadataa, joka mahdollistaa datan tulkinnan ja jatkokäytön. Tutkimusdatan metadata jaetaan ns. tutkimusprojektitason kuvaukseen (*study-level documentation*) ja datatason kuvaukseen (*data level documentation*). Tutkimusprojektitason kuvaus tukee löydettävyyttä ja sisältää tietoja mm. datan tuottajista/kerääjistä, datasetin nimen, julkaisuajankohdan, lisenssin jne. Tutkimusprojektitasolla kuvataan myös datan formaatti, tiedostorakenne ja versiointi, ja se sisältää tai siihen linkittyy kuvaus käytetyistä menetelmistä ja ohjelmistoista. Datatason kuvaus tarvitaan datan ymmärtämiseksi. Datatasolla kuvataan mm. tutkimuksessa käytetyt muuttujat. (Ks. *Document your data*, UK Data Service)

Datanhallintasuunnitelman tekeminen on usein ensimmäinen konkreettinen askel datan julkaisemisessa. Suomessa yleisesti käytetty suunnitelmapohja

(*General Finnish DMP template*) ohjaa suunnitelman laatijaa pohtimaan mm. käyttämänsä datan alkuperää, käyttöoikeuksia ja laatua sekä tuottamansa datan tallennusta, tietosuojaa, tarvittavia sopimuksia, dokumentointia, julkaisemista ja pitkäaikaissaatavuutta. Suunnitelmapohjan avulla tutkija tulee ajatelleeksi niitä asioita, joiden on oltava kunnossa, jotta datatuotos voidaan julkaista.

Tutkimusdatan vertaisarviointi ja laadun varmistus

Yleisesti ajatellaan, että tutkimusdata vaatii vertaisarviointiprosessin, jos sen halutaan olevan perinteiseen tutkimusjulkaisuun verrattava tutkimustuotos (Callaghan & al. 2012). Tutkimusdatan vertaisarviointi ei kuitenkaan ole täysin identtistä perinteisten julkaisujen vertaisarvioinnille. Tieteellisten seurain valtuuskunta (TSV 2014) määrittelee vertaisarvioituja tiedejulkaisuja koskevan tunnuksen ohjeistuksissa, että vertaisarvioinnissa tarkastellaan "aineiston kattavuutta ja teoreettisen viitekehyksen hallintaa, tutkimuksen toteutuksen luotettavuutta ja tarkkuutta sekä tulosten omaperäisyyttä ja uutuusarvoa suhteessa aiempaan tutkimukseen tieteenalalle ominaisella tavalla." Tutkimusdatan osalta arvioinnissa näkökulma on toinen ja siinä keskitytäänkin kuratointiin, arkistointiin ja datan laatuun (Parsons & Fox 2013).

Tutkimusdatan vertaisarviointia toki tapahtuu osana perinteisen tutkimusjulkaisun vertaisarviointia. Tiedekustantajat voivat vaatia vertaisarvioijia pyytämään tarkasteltavaksi myös datan, jonka pohjalle tutkimuksen tulokset rakentuvat (esim. Springer-Nature, *Research data policy type 4*). Tämä on vielä kuitenkin suhteellisen harvinaista. Varsinaiset datalehdet taas keskittyvät datan julkaisemiseen ja osa näistä lehdistä on vertaisarvioituja (esim. MDPI:n *Data*). Datalehdissä vertaisarviointi ei merkittävästi poikkea perinteisten julkaisujen arvioinnista, mutta toki sekin keskittyy datan laadun lisäksi sen teknisiin, käytettävyyteen ja saatavuuteen liittyviin näkökulmiin. Tarkemmin tutkimusdatan vertaisarvioinnin ohjeita on kartoittanut ja kuvannut Carpenter artikkelissaan (2017) *What Constitutes Peer Review of Data: A survey of published peer review guidelines*. Suomessa tiedelehti Terra otti jo vuonna 2014 yhdeksi artikkelityypiksi datankuvausartikkelin (*Data descriptions*), ja ensimmäisen datankuvausartikkelin julkaisemisen yhteydessä kertoi julkaisuprosessin noudattaneen normaalin vertaisarvioinnin vaiheita, ja että se paransi merkittävästi sekä käsikirjoitusta että dataa (Toivonen & Minoia 2014). Meillä on sittemmin myös laadittu ohje tieteellisille julkaisukanaville vastuullisen aineisto- ja datapolitiikan laatimiseksi (Ilva & al. 2020).

Perinteisten tieteellisten lehtien on hyvä ottaa huomioon tutkimusdata sekä artikkelien valinnassa, refereeohjeissa ja viittausohjeissa (Lilja 2017).

Tutkimusdataan liittyy käsite kuratointi. Data-arkistoiden tai -repositorioiden tekemään kuratointiin kuuluu laadunvarmistuksen näkökulma. Tutkimusdatan kohdalla vertaisarviointi voidaankin rinnastaan kuratointiin. Kuratoinnilla varmistetaan mm. datan dokumentoinnin ja metadatan laatu, formaattien käytettävyys ja datan rakenteen selkeys. Jokainen palvelu määrittelee omalle kuratoinnilleen tavoitetason. FAIR-periaatteiden (2016) tultua tunnetuiksi, nämä periaatteet ovat alkaneet osaltaan luoda tavoitetasoa. Moni data-arkisto ja -repositorio tavoitteleekin koneluettavuutta ainakin tutkimusdatan tutkimusprojektitason metatietojen osalta.

Kotimainen esimerkki kuratoidusta data-arkistosta on Tietoarkisto, FSD. Senkin kuratointi keskittyy tekniseen laatuun ja datan dokumentointiin. Käytettävyyden lisäksi varmistetaan arkistointikelpoisuus myös oikeuksien osalta (ks. FSD, Toimintaperiaatteet). Kansainvälisistä kuratoiduista data-arkistoista tunnetuimpia on The Reference Sequence (RefSeq), joka on DNA-, RNA- ja proteiinisekvenssejä sisältävä data-arkisto. Kuten usein suurien datamäärien kuratoinnissa, RefSeqin prosessi on kaksivaiheinen, jossa ensin käydään läpi automaattiset tietokoneen tekemät tarkastukset ja sen jälkeen laatua varmistetaan asiantuntijoiden työnä. RefSeqin kuratointityöhön osallistuu laaja kansainvälinen yhteistyöverkosto. Data tulee tietokantaan saataville nopeasti ja sen laatu paranee sitä mukaa kun tieto kyseisestä sekvenssistä karttuu. Datan kuratoinnin senhetkinen status ilmaistaan käyttäjille selkein merkinnöin.

Kaiken tyyppisille tutkimusdatoille ei ole olemassa omaa alanmukaista data-arkistoa, joka huolehtisi arvokkaiden data-aineistojen pitkäaikaissaatavuudesta. Arvokkaiden tutkimusdatojen säilymisen ja jatkokäytön turvaamiseksi Tieteen tietotekniikan keskus Oy (CSC) on rakentanut Fairdata PAS -palvelun. Se on kuratoitu palvelu, joka palvelee korkeakouluja. Fairdata PAS -palvelun kuratointiprosessi on hajautettu suurelta osin korkeakoulujen tehtäväksi. Korkeakoulut arvioivat data-aineiston arvon, varmistavat dokumentaation laadun ja täydentävät metatietoja, sekä sopivat datan käyttöehdoista. Palvelun rahoittava opetus ja kulttuuriministeriö varsinaisesti hyväksyy (tai hylkää) data-aineiston säilytettäväksi korkeakoulun hakemuksesta. Kuratointiprosessissa tehdään myös teknisiä tarkistuksia mm. tiedostomuotojen ja käytettyjen merkistöjen osalta. (Ks. tarkemmin CSC.)

Monet tiedekustantajat huolehtivat datan laadun varmistuksesta suosittelemalla julkaisuun liittyvien datojen julkaisemista luotettavissa data-arkistoissa ja -repositorioissa (esim. Springer Nature, *Research data policy type 1–3*). Osa näistä tietokannoista ei ole kuratoituja, mutta ne ovat kuiten-

kin laajasti tunnettuja ja paljon käytettyjä. Tutkimusdata, jota ei ole julkaistu vertaisarvioidussa datalehdessä tai kuratoidussa data-arkistossa, voi osoittaa laatunsa myös olemalla paljon käytetty (Kratz & Strasser 2015). Itsenäiset datasetit ja -tietokannat, joita tutkimusprojektit tai -infrat tuottavat, voivat siis osoittaa laatunsa tekemällä näkyväksi datan käytön. Tämä ei kuitenkaan ole käytännössä aina aivan yksinkertaista. Tässä palataan takaisin dataan viittaamiseen ja viitattavuuteen. Jos dataan on mahdollista viitata, myös viittausten seuraaminen helpottuu. Esim. DOI-tunnisteiden käytön seurantaa on rakennettu monia altmetriikkatyökaluja (mm. PlumX, Altmetrics), joiden avulla myös dataviittauksia voidaan seurata. Pysyviä tunnisteita siis tarvitaan myös datan laadun ja vaikuttavuuden osoittamisessa. Tämä onkin johtanut datan DOI-tunnisteiden käytön nopeaan lisääntymiseen (Hartgerink & Laakso 2020).

Miksi dataa julkaistaan?

Datan avaaminen ei välttämättä merkitse datan julkaisemista. Julkaisemisen näkökulma pitää sisällään yhteisesti hyväksyttyjen käytänteiden noudattamisen, kuten dataan viitattavuuden. Perinteisempien julkaisujen osalta keskeinen määrittelevä taho Suomessa on opetus- ja kulttuuriministeriö (OKM), jolle suomalaiset tutkimusorganisaatiot myös raportoivat osana perusrahoitusmalliaan tuottamansa, tietyt kriteerit täyttävät, julkaisut. Näissä kriteereissä ei toistaiseksi ole huomioitu tutkimusdataa, vaikka julkaistuna data on omiaan lisäämään tutkimuksen vaikuttavuutta, mikä näkyy mm. viittauksina (Parsons & al. 2010, Colavizza & al. 2020). Tutkimusrahoittajat ja tiedeyhteisö sen sijaan ovat jo mieltäneet tutkimusdatan tutkimusjulkaisun veroiseksi ansioksi, mikä näkyy eksplisiittisesti useissa yhteyksissä.

DORA-julistuksessa tunnustetaan tieteellisen tutkimuksen tuotosten moninaisuus ja erilaisuus, ja sellaisiksi mainitaan mm. uutta tietoa raportoivat tutkimusartikkelit, data, reagenssit ja ohjelmistot; immateriaaliomaisuus; sekä korkeasti koulutetut nuoret tieteentekijät. Tutkimusrahoittajia ja tutkimusorganisaatioita ohjeistetaan ottamaan tutkimuksen arvioinnissa tutkimusjulkaisujen lisäksi huomioon kaikkien tutkimustuotosten (sisältäen datasetit ja ohjelmistot) arvo ja vaikuttavuus. (Nykyri 2018.)

Tutkimuseettisen neuvottelukunnan (Tutkijan ansioluettelomalli. Tutkimuseettisen neuvottelukunnan suositus 2020) tutkijan ansioluettelomallin uusimmassa versiossa (2020, vrt. aiempi 2012) puhutaan aiempaa painokkaammin tutkimustuotoksista, mutta edelleen kehotetaan toimittamaan OKM:n luokittelun mukainen julkaisuluettelo erillisenä liitteenä.

Tutkimusaineistot ovat lisäksi esillä kohdassa “14. Tieteellinen ja yhteiskunnallinen vaikuttavuus”, jonka alla on mm. avoimen tieteen ja tutkimuksen edistäminen, esim. tutkimus- ja tietoaineistojen tuottaminen ja vastuullinen jakaminen ja tutkimustuotosten hyödyntäminen (sekä omien että muiden). Aiemmassa versiossa (2012) tutkimusdata-osuus oli esitetty kohdan “Tutkimustyön tieteellinen ja yhteiskunnallinen vaikuttavuus” yhteydessä, jonka alla oli “ansiot tutkimus- ja tietoaineistojen tuottamisessa ja jakamisessa”.

Monet tutkimusinfrastruktuurit tuottavat tutkimusdataa. Tutkimusinfrastruktuurien merkitys datan avaamisessa tunnistetaan entistä paremmin. Tutkimusinfrastruktuurit tuottavat datajulkaisuja itse, tai varsinaisen julkaisemisen tekevät infrastruktuuria käyttävät tutkijat. Molemmissa tapauksissa infran tekemä pohjatyö vaikuttaa merkittävästi datajulkaisun laatuun. Sen vuoksi monet uudet suositukset ja linjaukset koskevat juuri infrastruktuureja. Esimerkiksi OECD:n ja Science Europen julkaisema “*Optimising the operation and use of national research infrastructures*” (2020) kannustaa vahvasti infrastruktuureja datanhallinnan suunnitelmien tekemiseen sekä FAIR periaatteiden huomioimiseen. Suomen Akatemia pyytääkin infroilta datanhallintapolitiikan (DMPol). Julkaisun suosituksissa tutkimusinfrastruktuurien tuottaman datan käyttöä halutaan ymmärtää entistä paremmin ja infrastruktuurien toivotaan tuovan esiin datan käytön hyötyjä. Tästä syystä suositellaan viittausohjeita ja pysyvien tunnisteiden käyttöä. Käytännössä tämä onkin hyvä alku, mutta ei yksin vielä riitä kovin pitkälle, jos datan julkaisemista ja käyttöä ei seurata. Mutta kenen tehtävä on tutkimusinfratruktuurien tuottaman datan käytön seuranta?

Datan julkaisemisen yleistymistä on merkittävästi edistänyt se, että perinteistä tutkimusjulkaisua on yhä hankalampaa julkaista ilman dataa. Monet tiedekustantajat vaativat ns. “*Data availability statementin*” (Springer-Nature, Wiley, jne). Tässä lausunnossa julkaisun kirjoittajat joko kertovat mistä tutkimusdata on saatavilla tai selittävät mistä syystä dataa ei voida jakaa (esim. datan omistaa kolmas osapuoli). Tutkimusdataa julkaistaan myös menetelmien ja koodien julkaiseminen yhteydessä. Datan julkaisemista tehdään tieteen hyvien käytäntöjen ja tulosten verifioimisen näkökulmasta. Tieteen avoimuuden näkökulmasta mikään tutkimustuotostyyppi (menetelmä, koodi, data, tulokset kertova julkaisu) ei yksinään ole ylitse muiden. Sen sijaan mahdollisimman aukottoman kokonaisuuden julkaiseminen on tärkeää.

Kansallisessa suosituksessa tutkijanarvioinnin hyvistä käytännöistä (Vastuullisen tutkijanarvioinnin työryhmä 2020, s. 11) todetaan: “Keskeisin arviointikriteeri tieteelliselle tutkimukselle on sen tieteellinen, sisällöllinen laatu. Tutkijaa arvioidaan muodostamalla kokonaisarvio hänen toimintansa ja tuotostensa tieteellisestä laadusta.” Tutkimusdatan kontekstissa tutkija-

arvioinnin piirissä ovat tällä hetkellä (vertaisarvioiduissa) datalehdissä julkaistut datat, mutta nämä arvioidaan artikkeleina. Kansallisen tutkimusdatan avoimen saatavuuden osalinjauksen mukaan avoimen tieteen koordinaatio on laatimassa suosituksen hyvistä käytännöistä, kuinka tutkimusdataan liittyvä työ ja tutkimusdatan avaaminen huomioidaan tutkijan työssä ja kuinka siitä meritoidutaan (ks. Avoimen tieteen koordinaatio, Tieteellisten seurain valtuuskunta 2021).

Johtopäätökset

Hyvän datanhallinnan lähtökohtana on, että tunnistetaan tutkimuksessa kerättävä ja käytettävä tutkimusdata. Yliopistoissa erityisesti kirjastot tarjoavat tähän palveluita mm. oppaiden, ohjeiden, työpajojen, koulutuksen ja opetuksen muodossa. Tämä työ luo pohjaa tutkimusdatan julkaisemiselle.

Datanhallinnan palveluiden jatkokehittämisessä on tärkeää ottaa huomioon, miten tutkimuksen meritoitumiskäytänteet ja digitalisaatio etenevät. Tutkimusyhteisö on jo monilla tieteenaloilla kehittänyt toimivia ratkaisuja ja hyviä käytänteitä datanhallintaan ja julkaisemiseen liittyen. Kirjastojen ja datatuen tärkeä tehtävä on tunnistaa ne ja auttaa levittämään sekä tarvittaessa laajentamaan ja syventämään niitä. Etenkin suuremmissa yliopistoissa datatukea toteutetaan verkostomaisesti ja moniammatillisesti toimimalla, ja palvelussa hyödynnetään eri palveluyksiköissä sijaitsevaa osaamista. (Ks. toimintamalleista esim. Kuusniemi & al. 2021.) Palvelusuunnittelua tulisi tehdä kuitenkin tiiviissä yhteistyössä koko tiedeyhteisön kanssa, ei vain eri palveluiden kesken.

Verkostomaiseen moniammatilliseen toimintatapaan siirtyminen on ollut merkittävä edistysaskel. Nähdäksemme seuraava askel on nivoutua aiempaa kiinteämmin tutkimuksen tekoon ja olla kumppanina tutkijoille ja tutkimusinfrastruktuureille. Tämä edellyttää erityisesti yliopistokirjastojen tukipalveluroolien aiempaa suurempaa monimuotoisuutta ja mahdollisuutta edelleen erikoistua datanhallinnan asiantuntijana. Joissain kirjastoissa datanhallinnan tuen roolien lisäksi on panostettu vahvaan datatieteen osaamiseen (*data science*). Tästä malliesimerkki on National library of medicine (Yhdysvallat), jonka visio on ollut jo pidempään kehittää toimintaansa datatieteiden suuntaan (Fridsma, 2015). Varsin rikkaita roolitus- ja työnjakomalleja on kehitetty ja otettu käyttöön myös Alankomaissa (Jetten & al., 2021). Tampereen yliopiston kirjastossa on keväällä 2021 käynnistynyt ns. data manager -palveluiden tarjoaminen tutkimusprojekteille. Tähän on otettu mallia erityisesti Utrechtin yliopistosta ja Delftin teknillisestä yliopistosta. Data champion

-rooleja taas on ollut jo pidempään Aalto-yliopistossa (data agentit) ja Oulun yliopistossa (datanhallinnan asiantuntijat). Nykytilannetta kuvaa, että näille maailmalla jo hyvää vauhtia vakiintumassa oleville keskenään erilaisille rooleille ei löydy vielä suomenkielisiä vastineita. Tämä hankaloittaa merkittävästi aiheesta viestimistä. Roolien kehittäminen ja moninaistaminen on esillä kansallisessa tutkimusdatan avoimen saatavuuden osalinjausluonnoksessa (Avoimen tieteen koordinaatio, Tieteellisten seurain valtuuskunta, 2021) ja Unescon avoimen tieteen suositusluonnoksessa (UNESCO, 2020). Jotta roolien kehittäminen konkretisoituu, tulee tutkimusorganisaatioiden olla aktiivisia suositusten käytäntöön panemisessa ja valmiita mukautumaan tarvittaviin muutoksiin varsin nopeassakin aikataulussa.

Meillä on Suomessa kehitetty avoin ja ketterä kansallinen avoimen tieteen koordinaation malli, josta vastaa Tieteellisten seurain valtuuskunta opetus- ja kulttuuriministeriön rahoituksella. Kansallisessa datayhteistyössä on tähän asti korostunut erityisesti yliopistokirjastojen datanhallinnan asiantuntijoiden aktiivisuus suhteessa muihin tarvittaviin toimijoihin. Syynä lienee se, että kirjastot ovat osoittautuneet luonteviksi ja päteviksi tahoiksi vastaamaan tutkimusorganisaatioissaan esimerkiksi datapalvelun käynnistämisestä, konseptoinnista ja koordinoinnista laajassa yhteistyössä yli yksikkörajojen. Yhteis- ja jatkokehittäminen vaativat kuitenkin aiempaa enemmän myös laajemman tiedeyhteisön ja muiden palveluyksiköiden aiempaa merkittävästi aktiivisempaa osallistumista. Kansainvälisellä tasolla yhteistyö on tarvittavan laajemman tiedeyhteisön osalta kotimaista malliamme rikkaampaa esimerkiksi RDAn (Research Data Alliance) ja CODATAn (the Committee on Data of the International Science Council) piirissä.

Suomessa kehittämistyö voisi rikastua, jos datanhallinnan näkökulmasta keskeisten alojen (esim. informaatiotutkimus) tutkijoita osallistuisi aktiivisemmin myös kehityshankkeisiin ja ylipäänsä tällainen kehittämistyö mahdollistettaisiin joustavammilla rahoituskäytänteillä ja yhteistyömalleilla. Se voisi olla omiaan kirittämään data-aiheisten opetus- ja tutkimussisältöjen runsastumista ja yleistymistä.

Kun data nähdään tutkimusjulkaisuna, kohdistuu sen palvelutuottamiseen moninaisen osaamisen tarvetta. Yliopistokirjastot toteuttavat varsin ansiokkaasti perinteisen julkaisemisen tukea, mutta datan kohdalla tarvitaan sellaista erityistä osaamista, jota on vielä varsin harvalla. Tähän tulisikin panostaa voimallisesti niin alan koulutuksessa, osaamisen kehittämisessä työelämässä ja rekrytoinneissa. Asiat edistyvät tällä hetkellä pitkälti tutkimusrahoittajien vaatimukset edellä, mutta tutkimusorganisaatioissa tarvitaan laajempaa ja pitkäjänteisempää sitoutumista asian kehittämiseen. Kaukokatseisuutta tarvitaan myös, kun mietitään datan elinkaarta, jossa sen julkaiseminenkin voi

toteutua ja ajoittua paljon monimuotoisemmin kuin perinteisten tutkimusartikkeleiden ja datan pitkäaikaissaatavuudesta huolehtiminen edellyttää onnistuakseen varsin mittavia hoivapalveluita.

Palvelusuunnittelussa datan elinkaaren hahmottaminen ja elinkaaren eri vaiheet huomioon ottaminen sekä julkaisunäkökulman sisällyttäminen onkin tähdellistä. Palveluiden saattaminen ehyeksi vaatii hyvin laaja-alaista yhteistyötä ja lisäpanostuksia. Kansallisten ja kansainvälisten palveluiden ja infrastruktuurien rooli on tärkeä, mutta viime kädessä vastuu mahdollisista aukoista on usein tutkimusorganisaatiolla – joko ratkaisun edistäjänä kansallisella tai kansainvälisellä tasolla tai itse tai yhteistyössä tarvittavan palvelun toteuttaen. Tarvittavaa laajaa yhteistyötä edellyttävät esimerkiksi pysyvät tunnisteet datalle.

Arvokkaiden tutkimusdatojen olemassaolon tekee entistä näkyvämmäksi mm. Fairdata PAS -palvelu, johon kansallisesti arvokkaita data-aineistoja on alettu keräämään. Sinänsä erittäin tärkeän palvelun käytössä on kuitenkin valitettava pullonkaula. Käyttääkseen Fairdata PAS -palvelua, korkeakoulujen täytyisi pystyä kuratoimaan data-aineistonsa. Tämän johdosta jokaisen korkeakoulun on jollain tapaa hankittava kuratointiosaamista. Kuratointi vaatii tieteenalan tuntemusta, mutta sen lisäksi tarvitaan myös datanhallinnan erityisosaamista. Fairdata PAS -palvelun käyttöönotto vaatii tämän lisäksi juridista tukea datan käyttöehtojen sopimisessa ja tietosuojaan liittyvissä asioissa. Nähtäväksi jää kuinka nopeasti korkeakoulut kattavasti kykenevät pystyttämään tutkimusdatan kuratointipalveluita hyödyntääkseen Fairdata PAS -palvelua ja saadakseen sen hyödyt. Toistaiseksi useimmissa korkeakouluissa ei ole osoitettu tähän työhön tarvittavia resursseja, mikä on hyvin lyhytnäköistä. Näin toimimalla säästetään parin henkilötyövuoden verran vuodessa ja annetaan miljoonia maksaneiden tutkimusdatojen tuhoutua.

Hyvän datanhallinnan ja datan julkaisemisen hyödyt ovat paljon suuremmat kuin niiden vaatimat panostukset. Tämän osoittaminen konkreettisin mittarein ja luvuin ei kuitenkaan ole kovin yksinkertaista. Ilman selkeitä kannustimia ja ulkopuolista seurantaa harva organisaatio ryhtyy priorisoimaan datajulkaisemisen edellytyksiä. Onkin tärkeää kehittää mittaristoa, sillä mittaaminen vaikuttaa palveluiden sisältöihin ja resursoinnin kohdentamiseen. Edellytyksiin vaikuttavat myös kustantajien vaatimukset ja miten niissä suhtaudutaan aineistojen ja menetelmien vastuulliseen avoimuuteen ja saatavuuteen. Samalla kun kehitämme datan julkaisemisen edellytyksiä, on tärkeä kehittää myös arvioinnin avoimuutta ja mittareiden läpinäkyvyyttä laajassa vuoropuhelussa tutkimusyhteisön ja palveluntuottajien kanssa.

Lähteet

Avoimen tieteen koordinaatio, Tieteellisten seurain valtuuskunta (2021). (Toim. Nykyri, S., Päällysaho, S., Rosti, T. (pj.), Sunikka, A., Neuvonen, A., Kuusniemi, M. E.) *Tutkimusaineistojen ja -menetelmien avoimuus. Korkeakoulu- ja tutkimusyhteisön kansallinen linjaus ja toimenpideohjelma 2021–2025. Osalinjaus 1: Tutkimusdatan avoin saatavuus*. Vastuullisen tieteen julkaisusarja 5:2021, Tiedonjulkistamisen neuvottelukunta ja Tieteellisten seurain valtuuskunta. https://doi.org/10.23847/isbn.9789525995466

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., . . . Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC's environmental data centres. *The International Journal of Digital Curation, 7*(1). https://doi.org/10.2218/ijdc.v7i1.218

Carpenter, T. A. (2017). What Constitutes Peer Review of Data: A survey of published peer review guidelines. *arXiv*. https://arxiv.org/abs/1704.02236

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLoS ONE, 15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

CSC - Tieteen tietotekniikan keskus OY. https://www.csc.fi/fi/etusivu

Fridsma, D. B. (2015). A new vision for the National Library of Medicine. *Journal of the American Medical Informatics Association, 22*(5), 1111. https://doi.org/10.1093/jamia/ocv122

Given, L. M., & Olson, H. A. (2003). Knowledge organization in research: A conceptual model for organizing data. *Library & Information Science Research, 25*(2), 157–176. https://doi.org/10.1016/S0740-8188(03)00005-7

Hartgerink, C., & Laakso, M. (2020). Hypergraph in action: DOI Primer [Blogikirjoitus]. https://blog.libscie.org/doi-primer/

Hakala, J. (2017). E-viittaamisen ihanuus ja kurjuus. *Informaatiotutkimus, 36*(2). https://doi.org/10.23978/inf.65190

"Hallituksen esitys eduskunnalle avoimen datan direktiivin täytäntöönpanoa koskevaksi lainsäädännöksi," Lokakuu 2020. https://www.lausuntopalvelu.fi/FI/Proposal/DownloadProposalAttachment?attachmentId=12562

Ilva, J., Nykyri, S., Mustajoki, H., Parland-von Essen, J., & Syrjämäki, S. (2020). *Ohje tieteellisille julkaisukanaville vastuullisen aineisto- ja datapolitiikan laatimiseksi*. Vastuullisen tieteen julkaisusarja 4/2020. Tiedonjulkistamisen neuvottelukunta ja Tieteellisten seurain valtuuskunta. https://doi.org/10.23847/isbn.9789525995220

ISO 25964-1:2011ISO 25964 – the international standard for thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. http://www.niso.org/schemas/iso25964

Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M., van Gelder, C. W. G. (2021). Professionalising Data Stewardship in the Netherlands. Competences, Training and Education. Dutch Roadmap towards National Implementation of FAIR Data Stewardship (Version 0.1). *Zenodo*. https://doi.org/10.5281/ZENODO.4320505

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., . . . Wächter, J. (2006). Data publication in the open access initiative. *Data Science Journal, 5*, 79–83. http://doi.org/10.2481/dsj.5.79

Kratz J. E., & Strasser C. (2015). Researcher perspectives on publication and peer review of *data*. *PLoS ONE, 10*(2), e0117619. http://dx.doi.org/10.1371/journal.pone.0117619

Kuusniemi, M. E., Nykyri, S., Päällysaho, S., Rantasaari, J., Savolainen, E., & Sunikka, A. (2021). Datatukea rakentamassa – Katsaus koulutuksiin ja palveluihin. *Signum, 52*(4), 4–14. https://doi.org/10.25033/sig.101386

Lahtinen, A., Lukkarinen, A., Koivula, H., Liimatainen, J. O., Parland-von Essen, J., Tana, J., . . . Pääkkönen, T. (2020). Choosing and implementing persistent identifiers : Guide for research organisations. http://doi.org/10.5281/zenodo.4395767

Laine, H. (2018). Open science and codes of conduct on research integrity. *Informaatiotutkimus, 37*(4). https://doi.org/10.23978/inf.77414

Laine, H., & Nykyri, S. (2018). Dataviittaamisen tiekartta tutkijalle. *Informaatiotutkimus, 37*(2). https://doi.org/10.23978/inf.72999

Laine, H., Asmi, A., Bingham, E., Hakala, J., Laaksonen, H., Myllymäki, P., & Nykyri, S. (2018). *Tracing data: Data citation roadmap for Finland. Finnish Committee for Research Data*. http://urn.fi/URN:NBN:fi-fe201804106446

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation, 6*(2), 4–37. https://doi.org/10.2218/ijdc.v6i2.205

Lilja, J. (2017). Tieteelliset lehdet ja tutkimusdata. *Informaatiotutkimus, 36*(1). https://doi.org/10.23978/inf.63189

MDPI: *Data*, https://www.mdpi.com/journal/data, EISSN 2306-5729.

Nykyri, S. (2018). DORA-julistus (San Francisco Declaration on Research Assessment). [Suomennos] *Informaatiotutkimus, 37*(4). https://doi.org/10.23978/inf.77417

OECD/Science Europe (2020). Optimising the operation and use of national research infrastructures. *OECD Science, Technology and Industry Policy Papers*, No. 91. OECD Publishing. https://doi.org/10.1787/7cc876f7-en

Parland-von Essen, J., Fält, K., Maalick, Z., Alonen, M., & Gonzalez, E. (2018). Supporting FAIR data: categorization of research data as a tool in data management. *Informaatiotutkimus, 37*(4). https://doi.org/10.23978/inf.77419

Parsons, M. A., & Fox, P. A. (2013). Is Data Publication the Right Metaphor?. *Data Science Journal, 12*, WDS32–WDS46. http://doi.org/10.2481/dsj.WDS-042

Parsons, M. A., Duerr, R., & Minster, J.-B. (2010). Data Citation and Peer Review. *Eos, Transactions American Geophysical Union, 91*(34), 297–298. https://doi.org/10.1029/2010EO340001

The Reference Sequence (RefSeq). https://www.ncbi.nlm.nih.gov/refseq/about/

Springer Nature. Research Data Policy Types. https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096

Tietoarkisto (FSD). Toimintaperiaatteet. https://www.fsd.tuni.fi/fi/tietoarkisto/#toimintaperiaatteet

Tieteellisten seurain valtuuskunta, TSV (2014). Tunnus vertaisarvioidulle tiedejulkaisulle: Käytön edellytykset. https://www.tsv.fi/fi/palvelut/tunnus/kayton-edellytykset

Tieteen termipankki (10.5.2021). Avoin tiede:tutkimusaineisto. https://tieteentermipankki.fi/wiki/Avoin_tiede:tutkimusaineisto

Toivonen, T., & Minoia, P. (2014). Launching a new article type in Fennia: Data descriptions. *Fennia - International Journal of Geography, 192*(2), 79–80. https://fennia.journal.fi/article/view/48008

Tutkimuseettinen neuvottelukunta (TENK 2020). *Tutkijan ansioluettelomalli. Tutkimuseettisen neuvottelukunnan suositus 2020.* https://www.tenk.fi/sites/tenk.fi/files/TENKin_ansioluettelomalli_2020.pdf

UK Data Service (2021). Document your data. https://www.ukdataservice.ac.uk/manage-data/document.aspx

UNESCO. (2020). First draft of the UNESCO Recommendation on Open Science 2020. https://unesdoc.unesco.org/ark:/48223/pf0000374837

Vastuullisen tutkijanarvioinnin työryhmä (2020). *Tutkijanarvioinnin hyvät käytännöt. Vastuullisen tutkijanarvioinnin kansallinen suositus.* Vastuullisen tieteen julkaisusarja 5:2020. https://doi.org/10.23847/isbn.9789525995268

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G:, Axton, M., Baak, A. . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Yleinen suomalainen ontologia, YSO (2021). Tutkimusaineisto. http://www.yso.fi/onto/yso/p16752