KATSAUS

Tieteelliset lehdet ja tutkimusdata

Johanna Lilja
Tieteellisten seurain valtuuskunta
johanna.lilja@tsv.fi
http://orcid.org/0000-0002-3905-0035

This report summarises the papers and discussions presented at the Scholarly Journals and Research Data Seminar organised by the Federation of Finnish Learned Societies and the Finnish Association for Scholarly Publishing in February 2017. Stricter policies on storing research data in repositories and opening it are now being implemented. In fact, 27 per cent of research funders now require data archiving, including the Academy of Finland. The seminar brought together funders, researchers and representatives from journals and data archives to discuss how archiving and opening data should be carried out and the role played by journals. The questions asked included: Should journals require their authors to link their text to research data or should they only encourage such action? Should journals guide their authors to use central national or international data archives or should they establish their own separate data repositories, for example in connection with the Finnish national data service IDA?

Asiasanat: data; avoin tieto; tieteellinen julkaisutoiminta; tiedelehdet; tietoarkistot

Avoimen saatavuuden käsite liittyi pitkään lähinnä julkaisuihin. 1990-luvulla käynnistyneen Open access -liikkeen päämääränä oli, että kaikilla olisi vapaa pääsy tieteellisiin teksteihin (Open Society Institute, 2002). Vähitellen kuitenkin avoimuuden tavoite laajeni koskemaan tieteenteon koko prosessia. Ilmaisen pääsyn lisäksi avoimuuden kriteeriksi tuli myös vapaa jakaminen ja muokattavuus. Jonkinlaiseksi avoimuuden ideaaliksi hahmottui tilanne, jossa jo tutkimussuunnitelma on avoimesti arvioitavissa, tutkimusryhmät pitävät verkossa tutkimuspäiväkirjaa, jota muut tutkijat voivat kommentoida ja hyödyntää, tutkijoiden käyttämät ohjelmistot

Pysyvä osoite: https://doi.org/10.23978/inf.63189

perustuvat avoimeen lähdekoodiin, jonka avulla niitä kehitetään jatkuvasti, tutkimusdata avataan ja sitä uudelleenkäytetään, ja tutkimus julkaistaan avoimessa lehdessä tai rinnakkaistallennetaan avoimeen julkaisuarkistoon (ATT-hanke, 2015b; European Commission, 2012). Tällaisella avoimella tutkimusprosessilla on jo saatu lupaavia tuloksia ainakin malariatutkimuksessa (American Chemical Society, 2016).

Avoimen tieteen osa-alueista tutkimusdatan avoimuus on noussut julkaisujen ohella vahvimmin esiin. Datan tallentaminen ja avaaminen mahdollistaa sen uudelleenkäytön esim. vertailevassa tutkimuksessa tai aikasarjojen tuottamisessa. Aineistojen avaaminen parantaa myös tutkimuksen laatua, sillä kukapa haluaisi avata toisille epämääräiseltä näyttävää dataa. Opetus- ja kulttuuriministeriön käynnistämässä Avoin tiede ja tutkimus -hankkeessa on ollut useita projekteja ja työryhmiä datapalvelujen ja dataan liittyvän tiedonhallinnan kehittämiseen. ATT-hankkeen tavoitteena on, että uusista data-aineistoista on lisensoitu 25% vuonna 2017, 30% vuonna 2018 ja 50% vuonna 2020, ja niiden metatiedot löytyvät kansallisista metatietokatalogeista (ATT-hanke, 2015a). Datan avaaminen etenee myös rahoittajien vaatimuksesta. Sherpa/Juliet -palvelun tuoreen tilaston mukaan 27 % maailman tutkimusrahoittajista vaatii datan tallentamista arkistoon ja 12% suosittelee sitä (SHERPA/JULIET, ei pvm.).

Myös elinkeinoelämä on kiinnostunut datasta, joskin tässä data on käsitettävä tutkimusdataa laajemmin, esim. kuluttajiin liittyvänä tietona. Google-haku jo tutuksi tulleesta Clive Humbyn lausahduksesta *Data is new oil* toi esiin monia muitakin luonnehdintoja *Data is new currency, Data is the new bacon.* Tämä viimeisin, joka näyttää viittaavan pekoniin eikä Sir Francis Baconiin, on tuotteistettu jo t-paidoiksikin (Seiner, 2015; Toonders, 2014). Vaikka näitä lausahduksia nettikeskustelussa myös kyseenalaistettiin, kertovat ne kuitenkin omaa kieltään siitä, että tutkimusdatallakin on potentiaalista käyttöä myös tiedeyhteisön ulkopuolella.

Avoimen datan kaudella tieteellisiin lehtiin kohdistuu uusia odotuksia. Niiden pitäisi ottaa tutkimusdata huomioon sekä artikkelien valinnassa, refereeohjeissa ja viittausohjeissa. Maailmalla on syntynyt erityisiä datalehtiä *data journals* tai *open data journals,* joissa on datakuvauksia, metatietoja ja linkityksiä data-aineistoihin (Polydoratou, 2015). Suomessakin on otettu joitakin askeleita tähän suuntaan, esimerkkinä historian verkkolehti *Ennen ja nyt* ,joka pilotoi digitaalisten lähteiden linkittämistä PID-koodien avulla jo 2013 (Lahtinen, 2013). Maantieteelliset lehdet – suomenkielinen Terra ja kansainvälinen Fennia – ottivat datankuvausartikkelit uudeksi artikkelikategoriaksi 2014 (Toivonen & Minoia, 2014). Tiedejulkaisujen käytännöt ja vaatimukset tutkimusdataan linkittämisestä ovat kuitenkin – kansainvälisestikin katsottuna – edelleen hyvin kirjavia. Lehtien kannalta on elintärkeää, että ne pystyvät palvelemaan kirjoittaja- ja lukijakuntaansa avoimen tieteen edellytysten vaatimalla tavalla mm. ohjaamalla linkityksiin, joissa käytetään

pysyviä tunnisteita.

Tieteellisten seurain valtuuskunta, joka tarjoaa julkaisupalveluja jäsenseurojensa lehdille ja monografiasarjoille, sekä suomalaisia tiedekustantajia edustava Suomen tiedekustantajien liitto, päättivät syksyllä 2016 käynnistää keskustelun siitä, miten tieteellisten lehtien olisi hyvä muuttaa käytäntöjään jotta tutkimusdatan linkittäminen artikkeleihin vakiintuisi osaksi kotimaista tiedejulkaisemista. Tieteelliset lehdet ja tutkimusdata -seminaari järjestettiin 1.2.2017 ja sinne pyydettiin puheenvuorot Suomen Akatemialta tutkimusrahoittajan näkökulmasta, eri aloja edustavilta tutkijoilta ja lehdiltä sekä data-arkistoilta ja -palveluilta. Seminaariin osallistui 71 henkilöä, lehtien lisäksi myös kirjastojen henkilökuntaa.[1]

Tavoitteista käytäntöön

Johanna Lilja totesi päivän avauspuheessa, että vaikka datan tallentaminen on tavoitteellistettu ja vaikka datanhallintaa sujuvoittamaan on kehitetty työkaluja, tieteenteon arki perustuu kuitenkin monilla aloilla edelleen perinteiseen yksittäisten tutkijoiden tai tutkimusryhmien puurtamiseen, jonka tuloksiin pääsee tutustumaan vain tieteellisen julkaisun kautta. Kysymyksiä ja epätietoisuutta on vielä paljon. Hän viittasi omaan tutkijankoulutukseensa Tampereen yliopiston informaatiotutkimuksen alalla vuosina 2006–2011. Vaikka Yhteiskuntatieteellisen tietoarkiston palveluita jatko-opiskelijoille jo tuolloin esiteltiinkin, ei tutkijankoulutukseen kuulunut datanhallinnan käytännön taitojen opetusta. Hän arveli, että muutkin tutkijat kokevat vielä olevansa datanhallinnassa epämukavuusalueellaan erityisesti humanistisissa tieteissä, joissa dataa ei tuoteta itse. Tällä hetkellä koulutettavat jatko-opiskelijat jo todennäköisesti kasvavat datan avaamisen kulttuuriin, mutta vanhemmille tutkijoille nämä kysymykset eivät ole itsestään selviä.

Päivän esitykset aloitti Suomen Akatemian tiedeasiantuntija Jyrki Hakapää. Suomen Akatemia on vuodesta 2014 alkaen kehottanut tutkijoita tallentamaan datansa oman tieteenalansa kannalta tärkeään kansalliseen tai kansainväliseen arkistoon tai tallennuspalveluun ja avaamaan sen, jos se vain on mahdollista. Viime syksyn haussa rahoituksen hakijoilta edellytettiin ensimmäistä kertaa aineistonhallintasuunnitelman liittämistä hakemukseen. Suunnitelmassa kerrotaan, millaista tutkimusaineistoa hankkeessa syntyy, mihin se tallennetaan ja voidaanko se avata. Tutkijan vastuulla on arvioida avoimuuteen liittyvät mahdolliset rajoitukset ja esittää suunnitelmassa perustelut, jos aineistoa ei voi avata.

---

1 Päivän esitykset on linkitetty sivulle `https://www.tsv.fi/fi/koulutukset/tieteelliset-lehdet-ja-tutkimusdata-p%C3%A4iv%C3%A4n-esitykset-nyt-verkossa` Esitysten videotallenteet ovat katsottavissa 14.8.2017 asti.

Muissa aamupäivän esityksissä kuultiin eri tieteenalojen tutkijoiden näkemyksiä datan avaamisesta. Professori Timo Vesala käsitteli ilmakehään ja ekosysteemiin liittyvän datan keräämistä ja jakamista. Hänen kokemuksensa datan avaamisesta olivat pelkästään hyviä, sillä avaaminen oli useimmissa tapauksissa johtanut tiivistyvään kansainväliseen yhteistyöhön ja tarjonnut myös kirjoittajuuksia artikkeleissa, joissa omaa dataa oli hyödynnetty. Keskeinen huolenaihe tutkimusdatan tallentamisessa on kuitenkin metadatan tuottaminen, sillä sen laatu on vaihtelevaa ja joskus se unohtuu kokonaan. Tarvitaan selkeät, yleisiin metadatastandardeihin soveltuvat formaatit ja riittävää ohjeistusta tutkijoille metadatan tuottamisessa. Digitaaliset tunnisteet auttavat linkittämään datan sen tuottaneisiin tutkijoihin ja hälventävät siten tutkijoiden pelkoja tekijänoikeuden menettämisestä.

Kommenttipuheenvuorot Vesalan esitykseen saatiin mikrobiologian ja kuluttajatutkimuksen aloilta. Helsingin yliopiston Elintarvike- ja ympäristötieteiden laitoksen vastuullinen tutkija Sari Timonen kertoi, että geenitutkimuksessa julkaiseminen edellyttää sekvenssien tallentamista keskitettyihin tietokantoihin. Alalla on muutama suuri toimija, joiden tietojärjestelmät on sovitettu yhteen. Tietokantoihin on kehitetty niiden käyttöä helpottavia työkaluja. Timonen otti kantaa myös eettisiin kysymyksiin ja riskeihin, joita sekvenssidatan käyttöön liittyy. Dna paljastaa paljon esimerkiksi ihmisen terveystiedoista, joihin kohdistuva kiinnostus ei ole pelkästään tieteellistä. Sekvenssitietoa yhdistelemällä voidaan luoda myös tuhoisia asioita. Viime kädessä sekvenssitieto kuitenkin lisää ymmärrystämme luonnosta.

Ihmistieteitä edusti seminaarissa kuluttajatutkimus. Tällä alalla kerättävään dataan liittyy vahvasti henkilöiden tietosuoja, joka mm. määrittää, että tietoja saa käyttää vain siihen tarkoitukseen, joka kerättäessä ilmaistaan. Kuluttajatutkimuskeskuksen johtaja Päivi Timonen kertoi, että nämä rajoitukset kuitenkin poistuvat, jos data on anonymisoitu siinä määrin, ettei henkilöä voi tunnistaa. Tällöinkin on huolehdittava siitä, ettei tietoja yhdistelemällä saada selville tutkittavien henkilöllisyyttä.

Lehtien rooli ja arkistojen rooli

Lehtien roolia avoimen datan edistämisessä tarkasteltiin niin ikään luonnontieteiden ja ihmistieteiden näkökulmasta. Suomen Metsätieteellisen Seuran toiminnanjohtaja ja Silva Fennica -lehden toimitussihteeri Pekka Nygren esitteli neljä tasoa, joilla lehdet voivat määritellä suhteensa dataan. Tasoluokituksen on laatinut Virginian yliopiston piirissä toimiva Center for Open Science. Nollatasolla lehti on datan avaamisen ja dataan viittaamisen suhteen passiivinen tai enintään rohkaisee kirjoittajiaan tähän. Korkeimmalla eli kolmostasolla lehti ei julkaise artikkeleita, jollei

dataa ei ole tallennettu luotettavaan arkistoon ja analyysejä toistettu ennen julkaisemista. Kolmostasolla vaaditaan myös täydellistä läpinäkyvyyttä tutkimukselle tutkimussuunnitelman tallentamisesta alkaen. Nygren totesi, että kolmas taso on vielä aika kaukana tieteenteon arkipäivästä. Välitasojen ratkaisut olivat käytännöllisempiä. Ykköstasolla lehden kuuluu ohjeistaa dataan viittaamiseen ja vaatia kirjoittajaa ilmoittamaan, onko data käytettävissä. Kakkostasolla datan tallentaminen luotettavaan arkistoon ja avaaminen silloin, kun se on mahdollista, on julkaisemisen ehto. Nygren korosti, että lehtien tehtävä ei ole ylläpitää omia data-arkistoja. Center for Open Sciencen tasoluokituskin perustuu siihen, että lehdet käyttävät luotettavia data-arkistoja.

Historiallisen Aikakauskirjan päätoimittaja professori Anu Lahtinen toi kommenttipuheenvuorossaan esiin humanistisen tutkimusdatan erityispiirteitä. Historian alalla data on usein kvalitatiivista, monikielistä ja kansallisesti tai alueellisesti rajattua. Toisin kuin luonnontieteissä, joissa tutkija tuottaa datan itse, historioitsijan aineistot ovat jo kaikkien käytettävissä kirjastoissa ja arkistoissa. Tutkija voi siis tallentaa ainoastaan alkuperäisaineistosta jalostettua dataa, kuten tilastoja tai tietokantoja. Avoimia kysymyksiä vielä ovat, mikä taho vastaa tämän datan säilyttämisestä ja miten siihen viitataan.

Iltapäivän viimeisessä osuudessa esiteltiin erilaisia data-arkistoja ja -palveluita. Tietoarkiston kehittämispäällikkö Arja Kuula-Luumi esitteli Tietoarkistoa, jonka palvelut ovat laajentuneet yhteiskuntatieteellisistä aineistoista kattamaan myös humanistien tutkimusaineistoja. Hän korosti, että monet keskeiset julkaisijat vaativat jo tutkimusdatan tallentamista johonkin tunnettuun data-arkistoon. Tietoarkisto ottaa vastaan monimuotoista aineistoa, tilastollisen datan lisäksi, myös mm. litterointeja, päiväkirjoja ja tutkijan digikuvaamia aineistoja. Tutkija saa arkistosta opastusta jo tutkimuksen alkuvaiheessa. Tietoarkiston palvelut ulottuvat anonymisoinnin tarkastukseen ja metadatan tuottamiseen suomeksi ja englanniksi. Palveluportaali Ailan kautta tutkijat pääsevät etsimään heitä kiinnostavaa dataa.

Jessica Parland-von Essen CSC:ltä esitteli Avoin tiede ja tutkimus -hankkeen puitteissa tuotettua tutkimusaineistojen säilytyspalvelua IDAa, johon Suomen korkeakoulujen ja Suomen Akatemian rahoittamat tutkijat voivat tallentaa dataansa tietyn kiintiön sallimissa rajoissa. IDA vastaa aineistojen pitkäaikaissäilytyksestä. Järjestelmä luo aineistolle pysyvän tunnisteen automaattisesti. CSC on tuottanut myös tutkimusaineistojen hakupalvelun Etsimen, joka mahdollistaa paitsi aineistojen etsimisen eri tietokannoista, kuten Tietoarkistosta ja Kielipankista, myös oman metadatan luomisen ja omien tunnisteiden liittämisen. AVAA- palvelu on julkaisualusta avoimille aineistoille. Parland-von Essen ehdotti, että myös tieteelliset lehdet neuvottelisivat mahdollisuudesta hankkia omaa säilytystilaa IDAsta. Metadatan tuottaminen Etsimen avulla on pääasiassa tutkijan tehtävä, mutta lehdet voisivat ohjeistaa tutkijat kuvailemaan ja tallentamaan aineistonsa.

Lopuksi kuultiin Ari Lukkarisen esitys EUDAT-palvelusta, joka mahdollistaa eri tieteenalojen tutkimusaineistojen varastoinnin, säilytyksen, julkaisemisen ja haun moninaisille toimijoille. EUDAT 2020 on 33 toimijan yhteinen palvelu. Suomesta sen kehittämiseen osallistuu CSC.

Keskustelu käyntiin

Tieteelliset lehdet ja tutkimusdata -teeman ympäriltä syntyi vilkasta keskustelua eri esitysten ja päivän päättäneen paneelin yhteydessä. Suomen tiedekustantajien liitolle esitettiin toivomus, että se ryhtyisi laatimaan kotimaisille lehdille ohjeita ja ideoita tutkimusaineistoihin liittyvistä menettelyistä Center for Open Sciencen tasoluokitusten pohjalta. Liiton sihteeri Raimo Parikka lupasi viedä tämän ehdotuksen liiton hallituksen käsittelyyn.

Keskusteltiin myös, olisiko järkevää toimia Jessica Parland-von Essenin ehdotuksen mukaisesti ja neuvotella opetus- ja kulttuuriministeriön kanssa mahdollisesta omasta tallennustilasta lehdille IDA-palvelussa, vai olisiko parempi, jos lehdet ohjaisivat kirjoittajiaan tallentamaan alakohtaisiin data-arkistoihin. Todettiin, että lehtien omat data-arkistot hajauttaisivat aineistot moniin pieniin erillisiin arkistoihin, jolloin datan löytäminen ja hyödyntäminen kärsii. Myös Suomen Akatemian datalinjaus näyttäisi edellyttävän kansallisten ja kansainvälisten data-arkistojen ja tallennuspalveluitten käyttöä. Kotimaisten lehtien taloudelliset resurssit ovat hyvin niukat ja ne toimivat nytkin pitkälti vapaaehtoistyön varassa. Oman data-arkiston ylläpito ja siihen mahdollisesti liittyvä vastuu metadatan laadusta ja aineiston löytyvyydestä vaatisi osaamista ja henkilöresursseja, joita lehdillä ei tällä hetkellä ole. Todennäköistä onkin, että ainakin humanistiset ja yhteiskuntatieteelliset lehdet tulevat mieluummin ohjaamaan kirjoittajiaan Tietoarkiston palveluihin.

Esitettiin myös ajatus, että kotimaiset tiedelehdet voisivat vaatia kirjoittajiaan tallentamaan aina tutkimusaineistonsa ja avaamaan sen, jos se vain on mahdollista. Nykytilanteessa, jossa datanhallinta ei kaikilla tutkimusaloilla ole vielä arkipäivää, vaatimus kasvattaisi huomattavasti riskiä kirjoittajakunnan menettämisestä. Todettiin, että on olennaista, että datanhallintaan liittyvät kysymykset tulevat ensin kiinteäksi osaksi yliopistotutkijoiden työtä ja osaamista. Lehtien yhteistyö edustamiensa tutkimusalojen opetushenkilökunnan kanssa on myös tärkeää. Datan avaamiseen voisi rohkaista tutkijoita esimerkiksi palkitsemalla ansiokkaimpia artikkeleja, joissa tutkimusaineistot on avattu. Lopuksi ehdotettiin, että edistyksellinen datapolitiikka voitaisiin ottaa myös yhdeksi tekijäksi JUFO-tasoa määriteltäessä. Koska JUFO:sta ei ollut osallistujia paikalla, keskustelu tästä kysymyksestä siirtyi seuraavaan tilaisuuteen.

Tieteellisten lehtien datapäivä tuotti ehkä enemmän kysymyksiä kuin vastauksia. Rohkaisevaa kuitenkin on, että edessä on useita polkuja, joita voi seurata ja ilmassa on paljon hyvää tahtoa datanhallinnan edistämiseksi.

Kiitän Pekka Nygreniä, joka suunnitteli ja järjesti kanssani Tieteelliset lehdet ja tutkimusdata -seminaarin sekä tarkasti ja kommentoi tätä seminaariraporttia ennen sen lähettämistä julkaistavaksi.

Kirjallisuus

American Chemical Society. (2016). ”Open science” paves new pathway to develop malaria drugs. https://www.sciencedaily.com/releases/2016/09/160915133104.htm

ATT-hanke. (2015a). Avointiede - ATT-tavoitteet. http://avointiede.fi/att-tavoitteet

ATT-hanke. (2015b). Avointiede - Mitä avoimuus on. http://avointiede.fi/mita-avoimuus-on

European Commission. (2012). Commission recommendation of 17.7.2012 on access to and preservation of scientific information. Brussels: European Commission. https://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

Lahtinen, A. (2013). Kartano ja kyläläiset. Sundholman omistajien ja lähiseudun asukkaiden omaisuuskiistoja 1500-luvun Kalannissa | Ennen ja nyt. *Ennen ja nyt*, (2). http://www.ennenjanyt.net/2013/09/kartano-ja-kylalaiset-sundholman-omistajien-ja-lahiseudun-asukkaiden-omaisuuskiistoja1500-luvun-kalannissa/

Open Society Institute. (2002). Budapest Open Access Initiative. http://www.budapestopenaccessinitiative.org/read

Polydoratou, P. (2015). Data Journals. EC Workshop on Alternative Open Access Publishing Models. https://ec.europa.eu/digital-single-market/en/news/save-date-12-oct-ec-workshop-alternative-open-access-publishing-models

Seiner, R. S. (2015). Data is the New Bacon. http://tdan.com/data-is-the-new-bacon/18796

SHERPA/JULIET. (ei pvm.). World funders by publications archiving policy type. http://www.sherpa.ac.uk/juliet/stats.php?la=en&mode=simple

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

Toonders, J. (2014). Data Is the New Oil of the Digital Economy. https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/