Jukka Rantasaari

Datanhallinnan merkitys, tutkijoiden osaaminen ja kirjaston rooli kulttuurinmuutoksessa

Tutkijat pitävät hyvää datanhallintaa eli datan järjestelmällistä käsittelyä erittäin tärkeänä datan eheyden, tutkimustulosten luotettavuuden ja tutkimuksen toistettavuuden kannalta. Silti monet arjen tutkimuskäytännöt eivät palvele tätä tavoitetta. Kulttuurinmuutokseen tarvitaan koulutusta, palveluja ja kannustimia. Akateemisilla kirjastoilla tiedonhallinnan ja tiedonlähteiden hallinnan ammattilaisina on yksi avainrooleista muutoksen johtamisessa.

Suurta osaa eri tieteenalojen tutkimuksista ei ole onnistuttu toistamaan, Anu Silfverberg kirjoittaa elokuun Long Play lehdessä[1]. Vuonna 2016 Nature-lehden kyselytutkimukseen vastanneista 1 576:stä eri tieteenalojen tutkijoista 52 prosenttia piti tieteen toistettavuuskriisiä vakavana[2]. Vastaajien mukaan toistettavuutta voidaan parhaiten edistää paremmilla käytännöillä, koulutuksella ja kannustimilla. Suunnitelmallinen ja hyvin dokumentoitu datanhallinta edistää tutkimusprosessin läpinäkyvyyttä, datan uudelleen käytettävyyttä ja tutkimuksen toistettavuutta. Mutta miten on osaamisen laita?

Mitä on data?

Data on

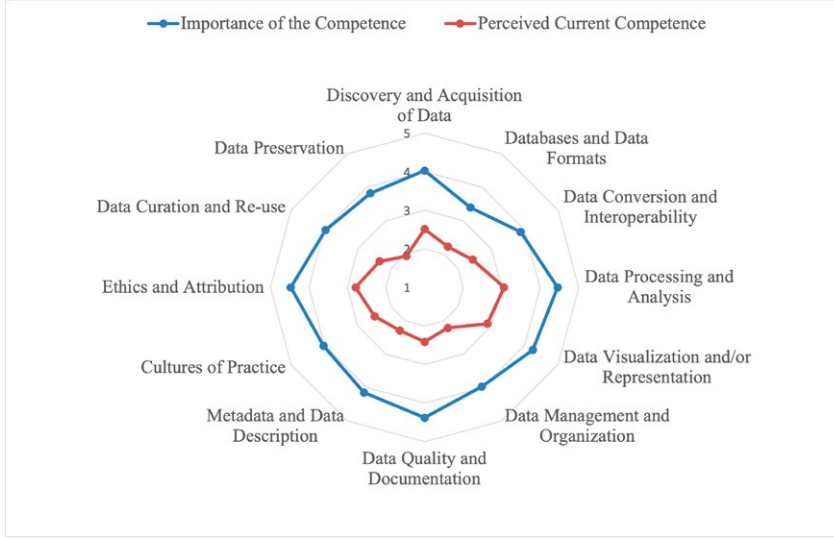
- kaikkea sitä mitä tutkija analysoi[3],
- kaikkea sitä informaatiota, jota tutkija systemaattisesti hankkii ja prosessoi uudeksi tiedoksi akateemisessa tutkimuksessa[4],
- keino validoida, evaluoida ja jäljittää se prosessi ja ne vaiheet, joilla tutkimustulokset on tuotettu[5].

Tutkimusdatan hallinta on

- datan systemaattista käsittelyä sisältäen toiminnot, joilla parannetaan datan löytyvyyttä, ymmärrettävyyttä ja käytettävyyttä nykyisessä ja tulevissa tutkimusprojekteissa[6].

Tutkimus väitöskirjatutkijoiden datanhallinnan osaamisesta ja osaamistarpeista

Haastattelin vuosina 2018–2019 Turun yliopiston eri tieteenalojen väitöskirjatutkijoita ja ohjaajia sekä lääketieteellisen tiedekunnan biostatistikoita (n=35). Haastatellut arvioivat datanhallinnan taidot keskimäärin erittäin tärkeiksi. Sen sijaan väitöskirjatutkijoiden tämänhetkinen osaaminen arvioitiin/itsearvioitiin keskimäärin tasolle ”osaa(n) jonkin verran” (kuva 1). Tämä ilmenee elokuussa 2021 julkaistusta tutkimuksestani[7].



Kuva 1: Akateemisten asiantuntijoiden (ohjaajat, biostatistikot) ja väitöskirjatutkijoiden arvioima datanhallinnan kompetenssien tärkeys ja samojen ryhmien arvioima/itsearvioima väitöskirjatutkijoiden tämänhetkinen osaaminen (1 = ei tärkeä/ei osaamista; 5 = välttämätön/erinomainen osaaminen; n=35).

No hyvä, näyttäisi että parantamisen varaa on. Mutta miksi se on niin tärkeää?

Monien tutkijoiden – rahoittajista puhumattakaan – pyrkimys on, että tutkimus ei hyödytä vain tutkijan urakehitystä vaan laajemmin tiedettä, yhteiskuntaa ja maailmaa[8]. Samaan tavoitteeseen pyritään tutkimuksen toistettavuuden ja luotettavuuden parantamisella. Lisäksi tutkimusyhteistyön lisääminen ja aikaisemman tutkimuksen tuotosten parempi hyödyntäminen mm. isojen maailmanlaajuisten ongelmien kuten ilmastonmuutoksen ja pandemioiden ratkaisemiseksi ovat perusteita datanhallinnan ja raportointikäytäntöjen

standardoimiselle sekä datan ja prosessin avaamiselle.

Tahdomme myös, että kansalaiset ja päätöksentekijät luottavat tutkimustietoon. Siispä meidän tulee lisätä ymmärrystä, miten olemme päätyneet tutkimustuloksiimme ja miten juuri tutkimustiedon tuottamisen tapa erottaa ne mielipiteistä, saduista, väärinkäsityksistä ja ennakkoluuloista.

Missä siis mättää?

Ongelma on, että monet arjen tutkimuskäytännöt eivät palvele näitä tavoitteita. Vaikka nuorilla tutkijoilla voi olla projekteissa iso vastuu tutkimuksen tärkeimmän raaka-aineen – datan – hankinnasta, tallennuksesta, prosessoinnista, analysoinnista ja säilytyksestä, he eivät yleensä ole saaneet datanhallinnan koulutusta, lukuun ottamatta tutkimusmenetelmäkoulutusta ja etiikan peruskurssia. Koulutuksen puuttuessa tutkijat selviytyvät ad hoc -ratkaisuilla ja yritys & erehdys -menetelmällä. Siksi käytännöt usein ovat epästandardeja, tutkimuksen immateriaali- ja sopimusasiat vieraita ja datan käsittelyvaiheet dokumentoitu meneillään olevaa tutkimusta ja sen tekijöitä, ei muita tutkijoita tai datan soveltajia silmällä pitäen. Tämä ei mahdollista datan jakamista ja uudelleenkäyttöä ja heikentää tutkimuksen toistettavuutta.

Suurimmat osaamiskuilut

Strukturoidun haastattelututkimukseni (Rantasaari, 2021) tulosten analyysi auttoi paikallistamaan suurimmat datanhallinnan osaamisvajeet neljään osa-alueeseen, joita olivat: (1) datan laatu, dokumentointi ja metadata, (2) datan jakaminen, uudelleenkäyttö ja pitkäaikaissäilytys, (3) datanhallinnan suunnittelu ja datan organisointi sekä (4) etiikka ja immateriaalioikeudet. Käyn seuraavassa läpi osaamisvajeet tarkemmin.

Datan laatu, dokumentointi ja metadata

”Documentation is made for myself to follow my data during the project.” (Doctoral student, Turku School of Economics).

Vaikka 80 prosenttia (n=12) väitöskirjatutkijoista uskoi dokumentointinsa olevan riittävän hyvää ulkopuolisenkin ymmärtää ja käyttää dataa, ainoastaan 15 prosenttia (n=3) haastatelluista akateemisista asiantuntijoista oli samaa mieltä. Datan dokumentointia ja kuvailua vaikeuttavat toisaalta datatyyppien moninaisuus, toisaalta se, että vaikka standardeja on jo luotu, niitä ei vielä tunneta ja käytetä.

”Most doctoral students don’t document because they don’t think anyone else would use their data after the current project” (Biostatistician, Faculty of Medicine)

Samaan aikaan kuin akateemiset asiantuntijat korostivat datan käsittelyn dokumentoinnin ja kuvailun tär-

keyttä, he tunnistivat koulutuksen ja standardien puutteen. Eräs ohjaaja lääketieteen piiristä totesi, että koska paineen parantaa datanhallintaa ja datan dokumentointia on koettu tulevan tieteenalan ulkopuolelta kuten päättäjiltä, tilastotieteilijöiltä ja data-analyytikoilta, dokumentoinnin taso on jäänyt matalaksi.

Datan jakaminen, uudelleenkäyttö ja pitkäaikaissäilytys

”In principle there has not been a thought that anyone other than researchers themselves would use their focus group interview data. As far as qualitative research data are concerned, there is not that kind of culture [of preserving data] as there is of preserving quantitative research data.” (Supervisor, Faculty of Social Sciences).

”Researchers’ focus is here and now, and they don’t pay so much attention to re-use and long-term preservation issues” (Supervisor, Faculty of Science and Engineering).

Haastattelemistani väitöskirjatutkijoista 87 prosenttia (n=13) suhtautui myönteisesti datan jakamiseen ja 73 prosenttia (n=11) arvioi datallaan olevan käyttöarvoa vähintään 50 vuoden ajan meneillään olevan tutkimuksen jälkeen. Silti he eivät tyypillisesti olleet huomioineet datan tulevaa käyttöä dokumentoinnissa eivätkä solmituissa datan käyttöä säätelevissä sopimuksissa – tai eivät olleet tietoisia tällaisten sopimusten olemassaolosta. Toisin sanoen, dataa ei mahdollisesti ollut lupaa jakaa ja käyttää meneillään olevan tutkimuksen jälkeen tai datan käsittelyprosessin dokumentoinnissa ja kuvailussa ei ollut huomioitu ulkopuolisia käyttäjiä. Myöskään nykyisten kannustinjärjestelmien ei koettu riittävästi tukevan jakamista ja uudelleenkäyttöä.

Vaikka kaikkea dataa ei ole mahdollista jakaa eettisistä ja juridisista syistä, tällöinkin usein voidaan jakaa metadata eli tutkimuksen ja datan kuvailutiedot. Lisäksi tutkimusluvalla, ns. ”pimeän arkiston” kautta on usein mahdollista jakaa rajoitetusti myös arkaluontoista aineistoa.

Datanhallinnan suunnittelu ja datan organisointi

”It would be important to have a big picture of the data and its relevance to understand the importance of preservation and re-use” (Supervisor, Faculty of Humanities).

”It would have been a huge benefit if there had been some training on data management” (Doctoral student, Faculty of Social Sciences).

Huolimatta siitä, että datanhallinnan suunnittelu ja datan hyvä organisointi arvioitiin erittäin tärkeiksi, asiaan on alettu kiinnittää huomiota vasta viime aikoina tutkimuksen digitoitumisen, datamäärien kasvun

ja yhteistyöprojektien lisääntymisen (e-research) myötä.

Etiikka ja immateriaalioikeudet

”Everything that has something to do with the letter of law is unclear and scary” (Supervisor, Faculty of Social Sciences).

Vaikka etiikka periaatteiden tasolla – johtuen ehkä pakollisesta etiikan kurssista – oli monien väitöskirjatutkijoiden itsearvioinnin perusteella jo verrattain hyvin hallinnassa, ohjaajat näkivät puutteita käytännöissä. Samoin käytännön toimenpiteitä kuvaavien vastausten perusteella lähes kaikki väitöskirjatutkijat kokivat epävarmuutta datan immateriaalioikeuksien, omistajuuden, sopimusten, tietosuojan ja käyttöoikeuksien suhteen.

Ei tässä syyllisiä kaivata vaan…

Jotta tutkimustulosten julkaisemisen lisäksi myös muiden tutkimustuotosten kuten datan, menetelmien ja koodin jakaminen ja hyödyntäminen yleistyisi, tarvitaan koulutusta, helppokäyttöisiä palveluja ja infrastruktuureja sekä laaja-alaisempia, tutkimuksen toistettavuutta edistäviä kannustimia[9], [10].

Kirjaston rooli

Kirjastot ovat sekä kansainvälisesti että Suomessa ottaneet johtavan roolin datanhallinnan palvelujen ja koulutusten suunnittelussa, koordinoinnissa ja tuottamisessa. Tämä on ymmärrettävää, koska data on volyymiltään ja merkitykseltään kasvava informaation lähde, ja kirjastojen ikiaikainen rooli on ollut kytkeä informaation lähteet ja informaation tarvitsijat yhteen. Tutkimusdatan hallinnan palvelujen lisääminen kirjaston palveluvalikoimaan merkitsee näkökulmien monipuolistamista: Pitkään kirjastoissa on keskitytty pääasiassa organisaation ulkopuolella tuotettujen tiedonlähteiden kuten painettujen ja sähköisten aineistojen valintaan, hankintaan, kuvailuun, järjestämiseen ja tiedonhankinnan opetukseen. Nyt myös tutkimusorganisaation omien tutkijoiden tuotoksilla kuten tutkimusdatalla, koodilla, ohjelmistoilla ja menetelmillä nähdään olevan tärkeä merkitys paitsi tutkijalle, laajemminkin tutkimukselle ja yhteiskunnalle. Tarvitaan palveluja, koulutusta ja resursseja myös näiden aineistojen valintaan, hankintaan, kuvailuun, järjestämiseen, tallentamiseen, pitkäaikaissäilytykseen ja jakamiseen[11].

Laadukas datanhallinta on monen eri alan asiantuntemusta vaativa kokonaisuus, johon tarvitaan kirjaston data-asiantuntijoiden lisäksi myös ainakin akateemisten asiantuntijoiden, tutkimuksen IT:n, lakiasioiden ja tietosuojavastaavan osaamista ja panosta. Tällä hetkellä datanhallinnan perusteiden koulutusta tarjotaan – yleensä kirjaston toimesta tai johdolla – jo monissa maamme korkeakouluissa. Koulutuksen laajuudet vaihtelevat yhden opetuskerran sessioista lukukauden toteutuksiin.

Turun yliopiston tutkijakoulussa on oma 3 ECTS Basics of Research Data Management (BRDM) -kurssi[12] väitöskirjatutkijoille ja postdoc-tutkijoille. Olemme kehittäneet ja järjestäneet kurssin vuodesta 2019 yhdessä eri tieteenalojen akateemisten asiantuntijoiden ja tutkimuksen tuen asiantuntijoiden kanssa, vuodesta 2020 yhdessä Åbo Akademin kanssa. Kurssi ei ole lopullinen ratkaisu datanhallinnan haltuun ottamiseen vaan pikemminkin ensi askel, joka auttaa nuorempia tutkijoita tunnistamaan puutteita nykyisissä datanhallinnan käytännöissä ja löytämään välineitä ja polkuja parempiin toimintatapoihin. Lisäksi, samalla kun osallistujat tutustuvat saatavilla oleviin tutkimuksen tukipalveluihin, tukipalvelujen asiantuntijat saavat tietoa tutkijoiden kohtaamista datanhallinnan käytännön haasteista (ks. kurssin rakenne ja osaamistavoitteet[13]). Valtakunnallisella tasolla Tieteellisten seurain valtuuskunnan (TSV) Datakoulutukset -työryhmä[14] kartoittaa ja kehittää osaamistavoitteita ja koulutuksia eri kohderyhmille.

Datanhallintaan liittyviä verkkokursseja

Data Carpentry[15]
"Introductory computational skills needed for data management and analysis in all domains of research."

Data Management Expert Guide[16]
"This guide is designed by European experts to help social science researchers make their research data Findable, Accessible, Interoperable and Reusable (FAIR)."

Data Management for Clinical Research[17]
"Critical concepts and practical methods to support planning, collection, storage, and dissemination of data in clinical research."

DataONE[18]
"n collaboration with the National Center for Ecological Synthesis and Analysis, DataONE has developed lessons, best practices, and training programs in data management to support research efficiency, productivity, and transparency."

Datatree[19]
"A free online course with all you need to know for research data management, along with ways to engage and share data with business, policymakers, media and the wider public."

Library Carpentry[20]
"Focuses on building software and data skills within library and information-related communities. Our goal is to empower people in these roles to use software and data in their own work and to become advocates for and train others in efficient, effective and reproducible data and software practices."

Mantra[21]
"MANTRA is a free online course for those who manage digital data as part of their research project."

Research Data Management and Sharing[22]
"Will provide learners with an introduction to research data management and sharing."

Linjaukset ja suositukset palveluiksi ja kannustimiksi

TSV:n Tutkimusaineistojen avoimen saatavuuden linjaus[23] julkaistiin tutkimusdatan osalta keväällä 2021. Tutkimusmenetelmien osalinjaus julkaistaan vuonna 2022. Keväällä 2022 julkaistavan Toimintakulttuurin avoimuuden linjauksen rinnalla julkaistaan tutkimusorganisaatioille suunnattu itsearviointityökalu mm. tutkimusdatanhallinnan palveluiden kehittämisen tueksi. Palvelujen itsearviointityökalun kriteerejä hyödynnetään myös valmisteilla olevassa Avoimen tieteen seurantamallissa.

Tarkista datanhallintasi

- Millä toimenpiteillä varmistat datan laadun kylmäketjun säilymisen läpi tutkimusprosessin? Esim. tarkistukset; tiedostojen versionhallinta, kansiorakenne ja nimeäminen; käyttöoikeuksien hallinta.
- Miten dokumentoit datan käsittelyn (valinnat, mittarit, työvaiheet) niin että sinä tai toiset pystyvät verifioimaan ja toistamaan käsittely- ja päättelyprosessin ja tarvittaessa uudelleenkäyttämään dataa? Esim. sähköinen laboratoriopäiväkirja; readme-tiedosto; kontrolloitu sanasto.
- Oletko sopinut tutkittavien ja projektin muiden tutkijoiden kanssa oikeudesta käyttää kerättyä, tuotettua ja anonymisoitua dataa myös meneillään olevan tutkimuksen jälkeen?
- Oletko selvittänyt datan omistus- ja käyttöoikeudet ja kirjannut keruutavan, rakenteen, muuttujat ja käyttöoikeudet esim. koodikirjaan, readme-tiedostoon tai muuhun oheis- eli metadataan?
- Onko data tallennettu tutkimuksen aikana riittävän tietosuojan ja varmuuskopioinnin takaaville turvallisille alustoille ja onko data tutkimuksen jälkeenkin saatavilla luotettavassa ja turvallisessa paikassa?

Jukka Rantasaari
Turun yliopiston kirjasto
jukka.rantasaari@utu.fi

Artikkelin kirjoittaja toimii kirjaston palvelupäällikkönä Turun yliopistossa ja valmistelee väitöskirjaa Åbo Akademiin tutkimusdatan hallinnasta. Rantasaari on myös mukana kansainvälisissä ja kansallisissa työryhmissä mm. valmistelemassa data-asiantuntijoiden osaamispolkua (RDA) sekä Toimintakulttuurin avoimuuden linjausta ja itsearviointityökalua (TSV).

Viitteet:

1. Silfverberg, A., 2021. Niin totta kuin osaamme. *Long Play*. 104. https://www.longplay.fi/jutut/niin-totta-kuin-osaamme

2. Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, (533), 452–454. https://www.nature.com/articles/533452a

3. Briney, K., 2015. Data management for researchers: Organize, maintain and share your data for research success. Exeter, UK: Pelagic Publishing.

4. Pryor, G., 2012. Why manage research data? In G. Pryor (Ed.), *Managing research data* (p. 224). Cambridge: Facet Publishing. doi:10.29085/9781856048910

5. Research Information Network 2008. Stewardship of digital research data: A framework of principles and guidelines.

6. Briney, 2015.

7. Rantasaari, J., 2021. Doctoral students' educational needs in research data management: Perceived importance and current competencies. *International Journal of Digital Curation,* 16(1), 1–36. https://doi.org/10.2218/ijdc.v16i1.684

8. Hurst, A., Pearce, A., Erickson, C., Parish, S., Vesty, L., Schnidman, A., Garlinghouse, M., & Pavela, A.,2016. *Purpose at work:* 2016 *global report.* https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/Global-Report-on-Purpose-at-Work.pdf

9. Chiarelli, A., Loffreda, L., & Johnson, R., 2021. Publishing reproducible research output. https://doi.org/10.5281/zenodo.4647697

10. Perrier, L., Blondal, E., & MacDonald, H., 2020. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. *PLoS ONE*, 15(2), 1–21. https://doi.org/10.1371/journal.pone.0229182

11. Ks. esim. Day, A., & Novak, J. 2019. The Subject specialist is dead. Long live the subject specialist! *Collection Management*, 44(2–4), 117–130.

12. Basics of research data management, 2021. Peppi Study Guide; University of Turku. https://opas.peppi.utu.fi/en/course/UGSL0001/13417

13. Rantasaari, J. et al. 2021. Basics of the research data management (brdm) course: Course structure and learning objectives 2019-22 | *Zenodo*. https://zenodo.org/record/5553794#.YYeZ_9ZBxuU

14. Tutkimusaineistojen avoimuus | Avoin tiede. (2021). Avoin Tiede; Tieteellisten Seurain valtuuskunta. https://avointiede.fi/fi/asiantuntijaryhmat/tutkimusaineistojen-avoimuus

15. *Data carpentry* 2021. Data Carpentry. https://datacarpentry.org/

16. *Data management expert guide – CESSDA TRAINING*. 2020. CESSDA. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide

17. Data Management for Clinical Research | Coursera, 2021. Coursera; Vanderbilt University. https://www.coursera.org/learn/clinical-data-management

18. Training | DataONE, 2021. Data Observation Network for Earth | DataONE. https://www.dataone.org/training/

19. Datatree – *Data training engaging end-users*, 2021. https://datatree.org.uk/

https://doi.org/10.1080/01462679.2019.1573708

20. Library carpentry, 2021. Carpentries. https://librarycarpentry.org/

21. Research data MANTRA. 2021. The University of Edinburgh. https://mantra.ed.ac.uk/

22. Research data management and sharing | Coursera. 2021, The University of North Carolina at Chapel Hill & The University of Edinburgh. https://www.coursera.org/learn/data-management

23. Tutkimusaineistojen ja -menetelmien avoimuuden linjaus | Avoin tiede. 2021. Avoin tiede; Tieteellisten Seurain Valtuuskunta. https://avointiede.fi/fi/linjaukset-ja-aineistot/kotimaiset-linjaukset/tutkimusaineistojen-ja-menetelmien-avoimuuden-linjaus