

Kooste miten tutkijat kuvailevat aineistojaan -työpajoista

- [Luonnontieteiden työpaja \(28.9.2022\)](#)
 - [Matti Ruuskanen, mikrobiologian ekologia, bioinformatiikka, mikrobiomitiede](#)
 - [Ari Mujunen, maa- ja ympäristötieteet, fysiikka ja tähtitiede](#)
 - [Pasi Kolari, ilmakehätiede, fysikaaliset tieteet, ekologia, evoluutiobiologia, informaatiotieteet](#)
- [Ihmistieteiden työpaja \(6.10.2022\)](#)
 - [Samuli Kaislaniemi, englantilainen filologia](#)
 - [Tero Alstola, Lähi-idän kulttuuri & kielet, kieliteknologia](#)

Luonnontieteiden työpaja (28.9.2022)

CSC ja Itä-Suomen yliopisto järjestivät yhteistyössä kaksi virtuaalityöpajaa, joissa kuulumme kuvailun käytännöistä tutkijoiden itsensä kertomana. Ensimmäinen työpaja keskittyi luonnontieteisiin ja esittelijöinä toimivat Matti Ruuskanen Turun yliopistosta aiheenaan mikrobiologia, Ari Mujunen Aalto-yliopistosta aiheenaan tähtitiede Metsähovin radiotutkimusasemalla ja Pasi Kolari Helsingin yliopistosta aiheenaan ilmakehätutkimus SMEAR-hankkeessa.

Matti Ruuskanen, mikrobiologian ekologia, bioinformatiikka, mikrobiomitiede

- [Matti Ruuskasen diat](#)

Tutkijatohtori Matti Ruuskanen kertoi mikrobiologian, bioinformatiikan ja datatieteen alojen käytännöistä oman tutkimustyönsä kautta. Tavoitteena on kuvailla tutkimusaineistoja siten, että koko tutkimus on toistettavissa projektin jälkeen ja FAIR-periaatteita noudattaen. Hän kertoi yksityiskohtaisesti miten data tuotetaan, säilytetään ja dokumentoidaan. Esimerkiksi on äärimmäisen tärkeää dokumentoida tarkka näytteenottomenetelmä, näytteiden käsittely sekä koodit, joilla raakadata on analysoitu, jotta tutkimus on mahdollista toistaa. Kuulijan näkökumasta dokumentointiprosessi näytti huippuunsa viritetyltä, eikä se jättänyt paljoa toivomisen varaa.

Ongelmiakin on tutkimusaineistojen kuvailussa tullut vastaan. Matti Ruuskanen kertoi toivovansa, että koulutuksessa kiinnitettäisiin enemmän huomiota datan ja analyysikoodin säilytykseen ja dokumentointiin ja että dokumentoinnin toteuttaminen tulisi ehkä ottaa jopa rahoitusehdoissa huomioon. Lisäksi hän kertoi, että vaikka standardointia on jo jonkin verran tehty ja siinä onnistuttu, etenkin fysiokemiallisen datan ja analyysikoodien osalta on vielä paljon kehittämistä. Sensitiivinen data tuo lisävaikeuksia: keräys ja dokumentointi on vielä fragmentoitunutta ja uudelleenkäyttäm mahdollisuudet rajattuja.

Ari Mujunen, maa- ja ympäristötieteet, fysiikka ja tähtitiede

- [Ari Mujusen diat](#)

Laboratorioinsinööri Ari Mujunen alustuksessa kuultiin, miten Aalto-yliopiston Metsähovin radiotutkimusasemalla mitataan auringon radiosignaaleja ja dokumentoidaan mittaustuloksia ja millaista suunnittelutyötä he olivat tehneet kuvailuun liittyen. He olivat tutkimusryhmässä pohtineet yhdessä, minkälaista kuvailua he tarvitsevat, sillä ajatuksella, että tuloksena olisi tiivis ja ytimekäs metadata. Varsinaisia standardeja ei ollut käytettävissä tällaiseen tarkoitukseen. Mujunen käyttää Qvain-työkalua metadatan kirjaamiseen ja kokee sen tarpeisiinsa sopivaksi työkaluksi. Hyötynä hän näkee myös pysyvät tunnisteet, jotka data saa ja joihin on helppo viitata. Heidän tuottamansa datan kohdalla on tärkeää, että käytössä on sellainen palvelu, joka mahdollistaa kertyvän aineiston kuvailun ilman että vanha data muuttuu. Tällöin uusi data tulee vanhan yhteyteen ja lisäksi. Data on myös löydettävissä [research.fi](#) -portaalista, ja muutkin integraatiot ovat mahdollisia.

Pasi Kolari, ilmakehätiede, fysikaaliset tieteet, ekologia, evoluutiobiologia, informaatiotieteet

- [Pasi Kolarin diat](#)

Pasi Kolari, Helsingin yliopiston Ilmakehätieteiden keskukselta kertoi miten [SMEAR](#)-asemilla toteutetaan datan kuvailua. SMEARin datavirrat on suunniteltu raakadatan keräyksestä ja prosessoinnista datan säilytykseen sekä jakamiseen ja julkaisuun loppukäyttäjille. Dokumentointi on kehittynyt huomattavasti hankkeen alkua ajoista, jolloin käytössä oli tekstimuotoisia päiväkirjoja ja tietoa dokumentoitiin pitkälti sisäpiirin ilmaisuin. Tänä päivänä hanke käyttää koneluettavia päiväkirjasovelluksia, jotka sisältävät muun muassa tiedot instrumenteista, joilla data on kerätty. He hyödyntävät myös verkkosivuja, joilla metatiedon kerääminen tutkimusryhmän kanssa yhdessä helpottuu. Koneluettavuus mahdollistaa myös tietojen nopean etsimisen.

Vaikeutena on ollut datan formaali kuvaaminen standardien avulla koneluettavasti. Tähän ei ole ollut tarjolla ohjeita eikä käytäntöjä, miten näitä voisi hyödyntää vapaamuotoisissa dokumenteissa. Huolimatta kuvailun haasteista pääajatuksena Kolari on kuitenkin pitänyt datan johdonmukaista kuvailua. Esimerkiksi julkaistaessa koodeja on tärkeää muistaa versiointi ja lisätä edes vapaamuotoinen teksti, jonka perusteella on mahdollista jäljittää, mitä datalle on tehty.

Ihmistieteiden työpaja (6.10.2022)

Toinen työpaja keskittyi humanistisiin tai ihmistieteisiin. Alkuperäisessä ohjelmassa oli neljä alustusta, mutta arkeologisen tutkimuksen kuvailuarjen asiantuntijat Mikko Suha ja Paula Kouki joutuivat kumpikin perumaan osallistumisen. Samuli Kaislaniemi ja Tero Alstola saivatkin ilman tiukkaa kelloon viikkuilua esitellä omien tutkimusalojensa kuvailukäytäntöjä ja omia kokemuksiaan kuvailun toteuttamisesta. Vilkkaalle keskustelulle jäi myös mukavasti aikaa.

Samuli Kaislaniemi, englantilainen filologia

- [Samuli Kaislaniemen diat](#)

Tutkijatohtori Samuli Kaislaniemi Itä-Suomen yliopistosta tutustutti yleisön englantilaisen filologian, tekstikorpusten ja sosiolingvistiikan maailmaan. Oman tutkimustaustansa esittelyssä hän sivusi myös tutkimusalan määrittelyn ongelmaa: Mihin kategoriaan rastiittaa tutkimuksen kuuluvan eri tiedonkeruun alustoissa, kun se ei taivu vain yhteen tutkimusalaan? Tätä taustaa vasten päästiin luontevasti kuvailun käytäntöjen paikoitellen hajanaiseen maailmaan.

Korpus on järjestelmällisesti kerätty, puhutun tai kirjoitetun kielen kokoelma, joka on kuratoitu ja koneluettavassa muodossa. Kokoelman muodostavia kielinäyteitä käsitellään jollain tavoin tutkimusta helpottamaan esimerkiksi muuttamalla fraktuurateksti nykyaakkoston mukaiseksi, jäsentämällä sanoja tai virkkeitä kielioipillisesti tai tekstin asettelua rakenteellisesti otsikoiden tai kappalejaon avulla. Korpuksesta voi myös olla useita eri formaatti- tai käsittelytavan versioita. Tämänkaltaiset tiedot ovat tutkijalle tärkeitä, jotta tarkasteltava aineistokokonaisuus asettuu kontekstiinsa ja jotta sen käyttö tutkimuksessa voidaan suunnitella ja toteuttaa. Kaislaniemen alustuksessa nousi erinomaisen konkreettisella tavalla esiin tutkijan näkökulma kuvailuun silloin, kun ollaan tekemässä tutkimusta ja kun tutkimuskysymys vaikuttaa merkittävästi aineiston yksityiskohtaisen kuvailun tarpeisiin. Esimerkiksi historiallisen sosiolingvistiikan kohdalla tarvitaan paljon metatietoa yksittäisten tekstien alkuperäisistä tuottajista ja käyttäjistä. Tällaista tietoa voivat olla vaikkapa sukupuoli, sääty, äidinkieli, koulutus, ikä, tekstin kirjoitusajankohta ja tarkoitus. Näitä ei kuitenkaan kysytä, kun kokonainen korpus julkaistaan. Korpusten kohdalla on huomionarvoista myös se, että niille ei ole olemassa yhtä yhtenäistä julkaisupaikkaa, ja siksi myös niiden metatiedot saattavat olla puutteellisia tai yksittäisiä korpuksia ei yksinkertaisesti tunneta, niitä ei löydetä ja sitä kautta niitä ei myöskään hyödynnetä.

Korpusten löydettävyyden ongelmaan ratkaisua on etsitty omin voimin. Kaislaniemen esimerkkinä on Helsingin yliopistossa toimivan VARIENG-tutkimusyksikön tuottama englannin kielen korpusten tietokanta CoRD (<https://varieng.helsinki.fi/CoRD/submissions.html#section4.2>). CoRD toimii korpusten hakukoneena, jossa eri korpukset kuvataan samalla metadatastandardilla. Sivusto ei sisällä itse korpuksia, mutta se tarjoaa niistä perustietoja. Ongelmallista on kuitenkin se, että kunkin korpuksen tiedon määrä vaihtelee, sillä korpusten takana olevat toimijat tuottavat tiedot itse.

Tämänkaltaisen metadatan tuottamiseen liittyy yleisemminkin ongelmia. Aina ei ole selvää, mitä eri metadatakenttiin tulisi kirjata ja miten tieto kuuluisi jäsentää, jotta data olisi löydettävissä ja sisällöllisesti ymmärrettävissä. Sama ongelma toistuu aineistohallintasuunnitelman kanssa: Tutkijalle ei välttämättä ole selvää, mitä eri kohtiin tulisi kirjoittaa tai ketä varten tietoa edes tarvitaan.

Tero Alstola, Lähi-idän kulttuuri & kielet, kieliteknologia

- [Tero Alstolan diat](#)

Tutkijatohtori Tero Alstola Helsingin yliopistosta kertoi Lähi-idän kulttuurin ja kielten tutkimuksesta sekä kieliteknologiasta erityisesti Muinaisen Lähi-idän imperiumit -projektissa (ANEE), jossa on noin 35 tutkijaa ja heistä lähes puolet ulkomailla. Aineistojen ymmärtäminen tutkijoiden kesken on siten tutkimuksen arkipäivää. Hyvänä taustatietona työpajalaisille Alstola kuvasi tiiviisti oman ryhmänsä tutkimuskysymyksen ja tutkimuksessa hyödynnettävät menetelmät. On tärkeä muistaa, että aineistot on tunnettava hyvin, jotta niitä voi lähteä digitaalisin keinoin tutkimaan tai tuottamaan. Ei riitä, että digitoidut tekstit ikään kuin sellaisinaan ovat verkossa tarjolla, vaan tekstejä on käsiteltävä eri tavoin. Esimerkiksi muinaisen Lähi-idän tutkimuksessa nuolenpääkirjoituksella kirjoitettujen savitaulujen kuvat ovat toki tarpeen ja hyödyllisiä, mutta tarvitaan myös esimerkiksi tekstien translitteraatiot, lukutapojen tulkinnat ja kielioipilliset jäsennykset. Alstola käytti esimerkkinä Oracc-korpusta (Open richly annotated cuneiform corpus), jossa eri projekteissa tuotettuja nuolenpääkirjoitustekstikokonaisuuksia on translitteroituina ja annotoituina, alan tiedeyhteisön yhteistyönä. Tämä korpus on viety Kielipankin KORP-palveluun ja kuvailtu Kielipankin kielivariantojen MetaShare-kuvailutietovarannossa, jossa on suhteellisen paljon koneluettavaa metadataa verrattuna vaikkapa Zenodon pakolliset kentät.

Tutkimusalojen perinteet tulivat tässäkin alustuksessa hyvin esille. Alalla on ollut tapana, että kerätystä datasta kirjoitetaan artikkeleit, mutta itse dataa ei ole julkaistu muiden käyttöön. Alstolan esimerkkitutkimuksessa on pyritty muuttamaan tätä perinnettä. Se on kuitenkin ollut vaikeaa, koska alalla ei ole esimerkiksi ollut selkeitä kuvailustandardeja käytössä. Kuvailu onkin jouduttu tekemään "tekemällä oppii"-periaatteella. Kantavana ajatuksena on pidetty kuvailun hyödyllisyyttä ja ymmärrettävyyttä loppukäyttäjille. Tämä tarkoittaa myös sitä, että tallennusalojen pakolliset metatietokentät eivät välttämättä riitä aineiston ymmärrettävyyden turvaamiseen. Vapaa kuvailu ja readme-tyyppiset tiedostot ovat tärkeitä ja joskus jopa helpompia tuottaa kuin pakolliset metatietokentät. Alstola painotti, että tutkimusryhmien kannattaa alusta asti käyttää aikaa keskeisten dokumentoitavien asioiden suunnitteluun ja systematisointiin. Tällöin tutkimusaineistoa voi hyödyntää tulevaisuudessa ja välttää tietojen penkomista projektin lopussa loputtomista sähköpostikeskusteluista. Vaikka tutkimusdatan ja metadatan tuottaminen ja datan julkaiseminen on arvokasta ja välttämätöntä alalle, Alstola muistutti myös äärimmäisen tärkeästä asiasta: tuo työ ei itsessään meritoi tutkijaa. Uudelleen käytettävän ja julkaistun datan tuottaminen koetaan helposti hukkaan heitetyksi ajaksi, koska aika on pois sellaisen tutkimuksen tekemiseltä, jolla tutkija voi meritoitua urallaan.