Hyvä aineistonhallinta lähtee ruohonjuuritasolta

Fakta Julkaistu 03.03.2016 klo 15:39 Kirjoittaja Suvi Pousi

Tutkimusaineistoilla tarkoitetaan niitä aineellisia tai aineettomia analysoitavia resursseja, joita tutkija tuottaa tai joita hän käyttää tutkimusprosessin aikana. Siihen sisältyy määritelmällisesti myös jonkinlaisia kuvailevia tietoja, eli vähintäänkin tieto aineiston alkuperästä ja sisällön laadusta. Datalla puolestaan tarkoitetaan aineiston kuvailematonta osaa, esimerkiksi empiirisin mittauksin saatuja lukuarvoja. Tutkimustyössä aineistonhallinnan suunnittelu on olennainen osa tutkimuksen suunnittelua. Tutkimusaineistojen tulisi olla käytettäviä, eheitä, virheettömiä, mahdollisuuksien mukaan avoimia ja luottamuksellisesti käsiteltyjä. Mitä tärkeämpi aineisto, sitä enemmän sen laatuun tulisi panostaa ja sitä huolellisemmin se tulisi dokumentoida ja kuvailla. Alasta riippuen tutkija voi myös meritoitua julkaisemalla aineistonsa avoimesti: standardimuotoisia aineistoja voidaan yhdistellä toisiin aineistoihin ja niitä voidaan käyttää uudelleen uusiin tutkimuksiin. Aineisto on myös tärkeä väline tutkimuksen luotettavuuden arviointiin, sillä aineiston avoimuuden arvioidaan ehkäisevän tutkimusvilppiä. Nykyään myös monet suuret rahoittajat vaativat rahoitushakemuksen liitteeksi aineistonhallintasuunnitelman, jolla tähdätään aineiston laadun varmistamiseen ja asianmukaiseen säilytykseen.

Jokainen tutkija haluaa pohjimmiltaan tuottaa mahdollisimman hyvälaatuista ja toistettavaa tutkimusta. Aineistonhallinnan rooli saattaa jäädä kuitenkin tutkimusprosessissa epäselväksi. Koko käsite voi olla vaikea mieltää: eikö aineistoja vain kerätä ja analysoida, miksi niitä pitää hallita? Aineistonhallinnan idean voi periaatteessa tiivistää hyvin lyhyesti: älä tee tyhmästi – mieti ensin.

Biostatistikko neuvoo: huolellinen pohjatyö säästää aikaa ja hermoja

Avoin tiede ja tutkimus -hankkeen osaajakoulutuksessa Turussa 11.2.2016 puhumassa ollut Turun yliopistossa työskentelevä biostatistikko Eliisa Löyttyniemi avaa esityksessään ansiokkaasti aineistonhallinnan eri puolia. Aiemmin lääketeollisuudessa työskennellyt Löyttyniemi kuvailee, miten yritysmaailmassa suuri osa tutkimuksen henkilöresursseista on nimenomaan suunnattu aineistonhallintaan. Lääkeyrityksissä lääkäri, statistikko, data management -henkilöt ja tutkimuksen koordinaattori miettivät yhdessä koko aineiston elinkaaren: mitä dataa kerätään, miten sitä kerätään ja mitä sille tapahtuu. Aineisto tarkistetaan huolellisesti joka käsittelyvaiheessa, jonka jälkeen se jäädytetään tilastollista analyysia varten eli siihen ei enää tehdä muutoksia. Lopullinen kerätty aineisto täyttää tutkimuksen tilaajan lisäksi sekä tutkimusta tekevän lääkärin että statistikon toiveet ja se on myös rakenteellisesti hyvin kerätty. Yliopistotutkijoilla ei yleensä ole käytössä tällaisia resursseja aineistonhallintaan, vaan tutkija vastaa yksin datan rakenteesta ja oikeellisuudesta. Tämä

voi johtaa ongelmiin: dataan voi päätyä virheitä, sen tulkinta voi olla hankalaa tai sen analysoiminen voi osoittautua vaikeaksi tai mahdottomaksi.

Löyttyniemen kuvaamat hyvän aineistonhallinnan vaiheet eivät ole rakettitiedettä – niitä ei vain välttämättä tule ajatelleeksi oikeaan aikaan. Hyvää aineistonhallintaa on esimerkiksi se, että päätetään *etukäteen* mitä muuttujia kerätään ja missä muodossa: kuinka monta numeroa tai desimaalia otetaan mukaan ja miten sanalliset muuttujat muotoillaan. Datalle asetetaan virheellisiä arvoja estäviä viiterajoja ja eri arvoja ristiintarkistetaan keskenään. Kaikista korjatuista arvoista kirjoitetaan ylös mitä arvoa on muutettu, miksi sitä on muutettu ja kuka sitä on muuttanut. Paperilta koneelle digitoitavat arvot ja ohjelmistosta toiseen siirrettävät arvot tarkistetaan oikeiksi. Alun perin tilastoanalyysin kannalta optimaaliseksi suunniteltu data on helppo ja hyvin nopea analysoida – tämä säästää tavattomasti tutkijan aikaa ja hermoja.


KEEP
CALM
AND
STUDY
BIOSTATISTICS

Eliisa Löyttyniemi puhumassa avoimen tieteen ja tutkimuksen osaajakoulutuksessa Turussa

Haasteina ohjelmistot, liian luovat ratkaisut ja inhimilliset virheet

Helppoa, eikö? Löyttyniemi kuvailee kuitenkin Turun lääketieteellisessä tiedekunnassa väitöskirjantekijöiden tekemiä virheitä, joista osa kuulostaa epämiellyttävän tutulta. Esimerkiksi tutkimukseen käytettävät kyselylomakkeet tulisi miettiä ja testata etukäteen – ei muuttaa niitä kesken tutkimuksen, vaikka niiden käyttö olisi todettu hankalaksi, käsitteet vaikeasti ymmärrettäviksi ja vastaukset moniselitteisiksi. Tutkimusaineistoa ei myöskään kannata kirjata niin luovaan muotoon, ettei sitä voi sellaisenaan siirtää tilasto-ohjelmaan. Ongelmia aiheuttavat varsinkin päivämäärät, värikoodaukset ja vaikkapa isojen ja pienten kirjaimien käyttö sekaisin jotakin asiaa koodattaessa. Huonosti suunnitellun ja käsitellyn aineiston korjaamiseen analyysia varten voi joutua käyttämään kymmeniä tunteja, puhumattakaan siitä minkälainen ajanhukka huonosti suunnitellun datasetin kokoamiseen liittyvä leikkaa-ja-liimaa-ruljanssi on.

Yksi suuri ongelma on Excelin ylivoimainen suosio tutkimusaineistojen hallinnassa – se kun ei ole tietokanta- vaan taulukkolaskentaohjelma. Tästä johtuen Excelissä pystyy saamaan datansa nopeasti sekaisin, ilman että sitä välttämättä lainkaan huomaa. Tietokantaohjelmien (esim. Access, Oracle) käyttö puolestaan vaatii tietokantaosaamista, jota ei opi hetkessä. Jos väitöskirjaohjaajallakaan ei tätä osaamista ole, ei ensimmäistä kertaa itsenäisesti aineistonhallintaa tekevä väitöskirjatutkija saa välttämättä mistään asiaan opastusta. Ilman tietokantaohjelmiakin voi kuitenkin selvitä, mutta Excelin käyttöön ja sen eri toimintojen opettelemiseen on suotavaa käyttää aikaa. Myös Exceliä käytettäessä datan oikeellisuuden tarkistukset, arvojen ristiintarkistukset ja tehtyjen korjausten ja muutosten dokumentointi on hyvin tärkeää. Lisäksi on mahdollista hyödyntää datan tallennusta suoraan tilasto-ohjelmaan (esim. SAS/JMP, SPSS), jolloin se on jo suoraan analysoitavassa muodossa.

Opetuksen oikea-aikaisuus tärkeää

Mistä tutkijoiden aineistonhallintaongelmat sitten johtuvat? Löyttyniemi mainitsee yhdeksi syyksi aloittelevien väitöskirjatutkijoiden kokemattomuuden ja tiedonpuutteen. Lisäksi Löyttyniemen mukaan syynä voi olla alan totuttu tapa tehdä aineistonhallintaa, sillä käytännöt periytyvät usein ohjaajalta ohjattavalle. Omasta kokemuksestani arvioisin, että tutkimuksen ollessa käynnissä tutkija voi kuvitella riittäväksi, että hän itse ymmärtää kerätyn datan ja tuntee sen käsittelyvaiheet. Tällainen ajattelu kuitenkin kostautuu usein analyysivaiheessa tai kun aineistoja pitäisi jakaa tutkimuskollegalle.

Aineistonhallinta sekä tilastotieteen perusosaaminen pitäisi periaatteessa sisäistää jo perustutkintovaiheessa. Joillakin aloilla tähän liittyvä opetus voi kuitenkin tulla käytännön tarpeiden kannalta väärään aikaan. Esimerkiksi Turun lääketieteellisessä tiedekunnassa biostatistiikkaa opetetaan Löyttyniemen mukaan jo toisena opintovuotena, jolloin se ehtii unohtua ennen väitöskirjan tekemistä. Löyttyniemi toivoisikin pitkin vuotta järjestettäviä tiiviitä syventäviä opintoja, joihin voisi osallistua silloin kun aineistonhallintaan liittyvät asiat tulevat ajankohtaisiksi. Väitöskirjaohjaajien roolia tulisi myös korostaa, sillä viime kädessä datakatastrofin estävät neuvot tulisi tulla heiltä. Lisäksi Löyttyniemen mielestä laitokselta tulisi löytyä aineistonhallinnan asiantuntija, jolta tutkijat voisivat kysyä neuvoa askarruttavista asioista.

Eri alojen aineistoilla on omat kompastuskivensä, mutta loppujen lopuksi hyvän aineistonhallinnan resepti on yksinkertainen: kysy neuvoa, suunnittele, tee muistiinpanoja, älä sählää. Näin tutkija säästää aikaa ja tuottaa laadukasta, analysointikelpoista ja tutkittavaa ilmiötä mahdollisimman hyvin

edustavaa dataa. Tutkijoiden onneksi aineistonhallinnan suunnitteluun on pian tulossa työkalu, DMPTuuli. Se ei hallinnoi aineistoja tutkijan puolesta, mutta sen avulla suunnitelmat eivät jää puolitiehen.

Eliisa Löyttyniemen videoitu esitys