DATAN MATKA INFORMAATIOKSI

REIJO SUND

Datan määrä tulee varmasti edelleen kasvamaan voimakkaasti. Tutkimuksen mielessä yhä isompi osa datan käytöstä tulee olemaan toisiokäyttöä eli data on alun perin tuotettu muuhun tarkoitukseen. Tarkoituksenmukaisen käytön mahdollistamiseksi pelkän datan avaamisen lisäksi on pystyttävä tuottamaan ja välittämään riittävästi taustatietoa ja ymmärrystä datan syntyprosessista ja käyttökelpoisuudesta, mikä tulee olemaan erittäin suuri haaste.

Mahdollisuuksia ja tulevaisuuden trendejä

Rekisteritutkimuksen mielessä erityisen kiinnostavaa on rekisteridatojen yhdistäminen muun tyyppisiin datoihin, johon henkilötunnus tarjoaa Suomessa kansainvälisesti katsoen poikkeuksellisen hyvät mahdollisuudet. Myös ylisukupolviset seurantatutkimukset, joissa henkilöistä ja heidän elinympäristöstään olisi monipuolisesti erilaista dataa koko elinkaaren ajalta, mahdollistaisivat uudenlaisia näkökulmia ja tutkimusasetelmia tutkimukselle. Todennäköisesti hyvin tunnettujen rutiiniluonteisesti kerättävien datojen hyödyntäminen tehostuu lähitulevaisuudessa, kun niistä saadaan puristettua reaaliajassa joihinkin määrättyihin tarkoituksiin hyödyllistä informaatiota. Arvelisin, että kuitenkin suurin osa myös isolla vaivalla avatuista datoista jää hyvin vähälle käytölle ja toisiokäyttö painottuu tiettyihin syystä tai toisesta kiinnostavaksi nousseihin datoihin. Uskon myös kunnollisen datan merkityksen korostuvan tutkimuksessa, mikä johtaa myös siihen, että korkeatasoisten julkaisukanavien vaatimukset sopivista datoista nousevat.

Isojen datojen käsittelyyn sopivien menetelmien yleistymisen myötä analyyseistä tulee kuvailevampia ja datakohtaisempia ja ne tulevat sisältämään enemmän subjektiivisiin preferensseihin perustuvia valintoja, jonka seurauksena on vaikeampaa tehdä datasta riippumattomia yleistyksiä. On myös odotettavaa, että datan käsitteen arkipäiväistymisen sekä roskadatojen ja -analyysien vuoksi usko tuotettujen tietojen luotettavuuteen käy läpi murroskauden. Tieteellisen tiedon erityisluonteisuuden säilyttäminen edellyttää tutkijoilta sen ymmärtämistä, että data on parhaimmillaankin vain hyvin kapea ja vääristynyt kuva todelli-

suudesta ja että siihen liittyvä epävarmuus ei johdu vain otannasta tai mittausvirheistä, minkä takia datan käsittely ja analysoiminen vaativat entistä enemmän taitoja ja tutkimukselle erityistä kriittistä otetta.

Rajoja ja rajoituksia datan käytössä

Merkittävimmät datan käytön rajoitukset liittyvät datan sisällölliseen antiin. Data ei ole sama asia kuin todellisuus vain parhaimmillaankin vain äärimmäisen kapea häivähdys jostain sen pienestä osasta. Käytännössä datan rajat määräytyvät sen käyttökelpoisuuden mukaan, joka vaihtelee tutkimuskysymyksestä toiseen, mutta joka usein puetaan sanoiksi laadun tai luotettavuuden termein. Iso osa datan käyttökelpoisuutta on dataan liittyvä taustatietämys. Mitä vähemmän taustatietämystä tarvitaan eli mitä enemmän datassa oleva perustuu suoraan havaittaviin asioihin, joista monilla ihmisillä on jo riittävän yhteinen käsitys, sitä suoraviivaisempaa dataa on käyttää. Välttämättömintä osaa taustatietämyksestä voidaan välittää myös niin sanotun metadatan avulla, mutta sellaisen tuottaminen erityisesti toisiokäytön osalta on haastava ja jatkuva prosessi. Teknisessä mielessä rajoituksia on kohdattu ja tullaan edelleen kohtaamaan datan koossa. Mikä on milloinkin liian paljon tai liian vähän riippuu tilanteesta ja on muuttunut nopeasti tekniikan kehityksen myötä. Joka tapauksessa on tietysti ongelmallista, jos dataa ei pystytä säilömään tai suorittamaan sille käsittely- tai analysointitoimenpiteitä kohtuullisessa ajassa.

Hiukan toisenlaisen näkökulman datan rajoihin ja rajoituksiin antavat eettiset pohdinnat, jotka liittyvät muun muassa siihen, milloin ja minkälaista dataa voidaan kerätä ja taltioida. Itse datan keräys sinänsä harvemmin vahingoittaa ei-kajoavana ainakaan ihmisiä, mutta yksityisyyteen ja tietosuojaan liittyvät kysymykset ovat tärkeitä niin datan säilömisen kuin varsinaisen käytönkin kannalta. Lainsäädännössä näihin asioihin on hiljattain otettu kantaa mm. GDPR:n ja SoTe-tietojen toisiokäytön lain myötä. Esimerkiksi SoTe-tietojen toisiokäytön laki pakottaa käyttämään lain alaisia tietoja vain tietoturvallisissa (etä)käyttöympäristöissä. Periaatteessa tämä mahdollistaa joissain tilanteissa sen, että myös tutkimuslupia vaativia (ainutlaatuisia suomalaisia rekistereistä peräisin olevia) sensitiivisiä datoja saadaan helpommin käyttöön. Kääntöpuolena on kuitenkin se, että lainsäädännön piiriin kuuluvien datojen yhdistäminen muuhun (tunnisteelliseen) tietoon todennäköisesti hankaloituu. On myös erikoista, että tutkijat pakotetaan lainsäädännöllä maksullisen tietoturvallisen käyttöympäristön käyttäjiksi eli siis asetetaan tiettyjen datojen käyttö maksumuurin taakse. On selvää, että tietoturvallisesta ympäristöstä ja datojen kokoamisesta aiheutuu ylläpitokustannuksia, mutta toivottavasti keskustelu suuntautuu siihen, löytyisikö kansallisia linjauksia, joilla edistettäisiin avointa tiedettä myös tällä saralla sen sijaan, että omatoimisesti vaikeutettaisiin juuri näiden Suomen vahvuuksiin kuuluvien ainutlaatuisten datojen käyttöä.

Kohti hyödynnettävää informaatiota

Keskeinen kriteeri muunnettaessa dataa hyödylliseksi informaatioksi on, että informaatio tuotetaan tieteellistä menetelmää käyttäen. Käytännössä kyse on siis empiirisestä tutkimuksesta ja yhä useammin käytössä on sekundaarista dataa. Peruslähtökohtana voidaan pitää sitä, että on olemassa todellisuus, jossa on ilmiöitä ja että näistä ilmiöistä voidaan tehdä jonkinlaisia havaintoja. Havainnot voidaan puolestaan operationalisoida siten, että ilmiöistä pystytään tekemään ”mittauksia”. Dataksi kutsutaan systemaattisesti tehtyjä mittauksia, jotka on tallennettu symboliseen muotoon. Kaikkea ei voida havaita ja valittujen havaintojen operationalisoinnit on mahdollista tehdä lukemattomilla tavoilla, joten parhaimmillaankin data on vain äärimmäisen kapea häivähdys todellisuudesta, jota sen ehkä voidaan kuvitella heijastelevan.

Kun kyseessä on datan toisiokäyttö, niin tutkijalla ei ole enää edes mahdollisuutta räätälöidä havaintoja ja operationalisointeja mahdollisimman tarkoituksenmukaisiksi tutkimuskysymyksen kannalta, vaan data on mitä on eikä sillä voida välttämättä vastata kaikkiin tutkimuskysymyksiin. Tämä tekee datan toisiokäytöstä helposti opportunistista: katsotaan sitä, mitä datasta voidaan nähdä, eikä sitä, mikä oikeasti olisi kiinnostavaa. On myös syytä pitää mielessä, että data on sitten vain tietyllä tavalla konstruoitu hiekkalaatikko, jossa voidaan vapaasti leikkiä hiekkalaatikon säännöillä. Kuiten-

kin vasta jos analyysien tuloksille pystytään antamaan perusteltu tulkinta eli suorittamaan niille operationalisoinnin käänteisoperaatio, voi data kertoa jotain hyödyllistä todellisuudesta. Tällaista empiirisen tutkimuksen tutkimusprosessia voidaan hahmotella eri tavoilla vaiheiksi, joilla data muuntuu informaatioksi ja samalla tällainen tutkimusprosessin kuvaus heijastelee hyvin myös datan matkaa informaatioksi.

Kirjallisuus

Sund, Reijo (2015). Miksi isoon dataan hukutaan? *Tieto & Trendit – Talous ja hyvinvointikatsaus* 2/2015, 40–45.

Sund, Reijo (2003). Utilisation of Administrative Registers Using Scientific Knowledge Discovery. *Intelligent Data Analysis* 7:6, 501–519.

Sund, Reijo, Gissler, Mika, Hakulinen, Timo ja Rosen, Måns (2014). Use of health registers. Teoksessa Ahrens, Wolfgang ja Pigeot, Iris (toim.): *Handbook of Epidemiology*, 2nd edition. Springer, New York, 707–730.

Sund, Reijo, Nylander, Olli ja Palonen, Tuula (2004). Raa'asta rekisteriaineistosta terveyspoliittisesti relevanttiin informaatioon. *Yhteiskuntapolitiikka* 69:4, 372–379.

Kirjoittaja on rekisteritutkimuksen professori Itä-Suomen yliopistossa.