Tilasto- ja tutkimusaineistojen avoimempi käyttö

Alustus seminaarissa "Tutkimus, aineistot ja avoimuuden rajat"

21.4.2008

Jussi Simpura

Tilasto- ja tutkimusaineistojen avoimempi käyttö

1. Otsikon tulkinta ja puhujan tausta
2. Avoimemman käytön rajoittajia
3. Keinoja avoimemman käytön mahdollistamiseksi
4. Haaveet ja todellisuus aineistojen käytössä

1.1. Tilasto- ja tutkimusaineistojen avoimempi käyttö: otsikon tulkinta

Tilastoaineisto:

Tilastoviranomaisen tai muun tilastoja laativan viranomaisen tilaston laatimista varten keräämä aineisto, jolla voisi olla myös tutkimuskäyttöä

Tutkimusaineisto:

Tutkimuslaitoksen tai siellä toimivan tutkimusryhmän keräämä aineisto, jolla voisi olla käyttö muissakin tutkimuksissa

1.2. Tilasto- ja tutkimusaineistojen avoimempi käyttö: puhujan tausta

Vuosikausia aineistojen kerääjänä, analysoijana ja käyttäjänä (erityisesti survey -aineistot)

Viisi vuotta Tilastokeskuksen tilastoeettisen lautakunnan jäsenenä (TK:n Elinolot -yksikön tilastojohtajana)

Mukana laatimassa lukuisia ehdotuksia TK:n aineistojen käytön helpottamiseksi

Tietoarkiston neuvottelukunnan jäsen 2008-

2. Tilasto- ja tutkimusaineistojen avoimemman käytön rajoittajia

(2.0. Vertailu Tietoarkiston listaan)

2.1. Tietosuojanäkökohdat

2.2. Tiedonomistusnäkökohdat

2.3. Tiedonkäyttötaitoihin liittyvät näkökohdat

2.4. Kustannusnäkökohdat

2.5. Tietorakennenäkökohdat

2.0. Tietoarkiston julkaisun (Borg & Kuula 2007) mainitsemia rajoittajia

Taitamattomuus aineistojen käytössä ja aineistojen soveltumattomuus uusiin tutkimusongelmiin

Epäselvyydet aineiston omistajuudesta

Akateeminen kilpailu

Aineistojen tietotekniset ja dokumentoinnin puutteet

Tutkimuseettiset tietosuojakysymykset

2.1.a. Tietosuojanäkökohdat rajoittajina: peruskysymyksiä

Aineistojen käyttötarkoitus: tilastoaineistot saatu usein "vain tilaston laatimista varten"

- joko tilastoviranomaisen oikeudella
- tai informoidun suostumuksen kautta

Myös tutkimusaineistoissa useimmiten informoitu suostumus

- lisärajoitteena mahdolliset yhdistetyt rekisteritiedot

Aineistojen käyttöön ja yhdistelyyn liittyvät tunnistamisvaarat

2.1.b Tietosuojanäkökohdat rajoittajina: lisäkysymyksiä

Tilasto- ja tutkimuseettisten toimikuntien vaihtelevasti perustellut käytännöt

- Merkkejä linjan jatkuvasta kiristymisestä?

Mitkä ovatkaan tietoarkistotyyppisten toimintojen eettiset säännöt?

Sekä tilasto- että tutkimusaineistot on usein kerätty vain yhtä ja tiedonantajalle erikseen ilmoitettua tarkoitusta varten. Tämä voi merkittävästi rajoittaa avoimempaa käyttöä.

2.2. Tiedonomistusnäkökohdat rajoittajina

Tavallinen vaatimus: "Verovaroilla kerätyt tiedot on saatava avoimempaan käyttöön" ("veronmaksajat" omistajina)

Kuka oikeastaan on aineiston omistaja?

- Tilastopuolella selvää: tilastontuottaja (myös asiakasrahoitteisesti kerätyissä aineistoissa)
- Tutkimuslaitospuolella epäselvempää: yksittäinen tutkija, tutkimusryhmä tai -hanke, tutkimuslaitos?
- Mitkä ovat omistajan oikeudet ja velvollisuudet?

2.3. Tiedonkäyttötaitoihin liittyvät näkökohdat rajoittajina

Kaikki aineistot ovat monimutkaisempia kuin kukaan käyttäjä voi etukäteen ymmärtää

Aineistojen tuottajilta vaaditaan todella tasokkaita metatietoja

Silti tapahtuu virhetulkintoja

=> Metatietojen lisäksi on saatava opastusta aineiston hyvin tuntevalta asiantuntijalta (ja ymmärrettävä pyytää tällaista, ja ymmärrettävä, että se voi maksaa)

2.4. Kustannusnäkökohdat rajoittajina

Yleinen harhaluulo: "Aineistojen kerääjät velkovat käyttäjiltä uudelleen kuluja, jotka on jo verovaroilla kerran maksettu" -> ei nähdä aineistojen irrotukseen ja ylläpitoon uppoavaa työtä.

Aineistojen kerääjillä ja käyttäjillä on toisistaan poikkeavia näkemyksiä siitä, mikä on kallista ja mikä ei

Tutkimusrahoituksen tyypillisiin volyymeihin nähden aineistokustannukset voivat olla korkeita

2.5. Tietorakennenäkökohdat rajoittajina

Varsinkin haastattelu- ja kyselytutkimusaineistot on usein rakennettu erityisiä kysymyksenasetteluja silmällä pitäen -> eivät taivu kaikkiin uusiin tehtäviin

Aineistoihin mahdollisesti liitetyt rekisteripohjaiset tiedot vaativat erityisiä lupamenettelyjä

Peräkkäisten saman aiheen tiedonkeruiden tietosisältö saattaa muuttua -> todellisten aikasarjaluonteisten aineistojen tuottaminen vaativaa!

3. Keinoja tilasto- ja tutkimusaineistojen avoimemman käytön mahdollistamiseksi

Ajankohtaista 1: Tilastokeskusta koskevan selvitysmiestyön yhtenä selvityskohteena aineistojen avoimempi käyttö

Ajankohtaista 2: Stakesin ja KTL:n fuusio antaa mahdollisuuden tarkastella laitosten tietovarantoja kokonaisuutena ja kehittää aineistojen avoimempaa käyttöä ainakin tulevan uuden laitoksen sisällä.

3a Keinoja viidellä rintamalla

3.1. Toimintasääntöjen selkiyttäminen

3.2. Teknisten ratkaisujen etsiminen

3.3. Tietovarantojen ylläpidon ja käytön tukemisen näkeminen infrastruktuuri-investointina

3.4. Tutkimusrahoituksen käytäntöjen muuttaminen: rahoitusta myös aineistojen hankintaan

3.5 Aineistojen käytön opastus paremmaksi

3.1. Toimintasääntöjen selkiyttäminen

On erittäin hyvä, että on muodostunut jo paljonkin säännöstöä tilasto- ja tutkimusaineistojen käytöstä (tilasto- ja tutkimuseettiset neuvottelukunnat, periaatteet ja ohjeistot)

Ilmeisesti ei kuitenkaan ole ollut mahdollista vielä paneutua näiden käytäntöjen ja ohjeistojen yhteensopivuuteen

Käytännössä eri suunnilta tietoja hankkiva tutkijataho joutuu tutustumaan kirjavaan sääntelykenttään

3.2. Teknisten ratkaisujen etsiminen

Tilastoaineiston tuottajan kannalta olisi ihanteellista, jos tekniset ratkaisut perustuisivat sopivasti supistettuihin ja tunnistamattomiksi tehtyihin aineistoihin, joita voisi etäkäyttää käyttölupamenettelyin.

(Vaihtoehtona kalliimpi laboratoriomalli)

Tutkijanäkökulmasta tuollaiset aineistot voivat olla liiaksi yksinkertaistettuja, ja tunnistamisvaaran välttely voi pakottaa hävittämään olennaista informaatiota.

3.3. Tietovarantojen ylläpito ja käyttö infrastruktuuri-investointina

>>> Tämä näkökulma pitäisi saattaa kaikkien tutkimusrahoituksesta päättävien laitosten ja tahojen tietoon

Useimmat tietoaineistojen kerääjät ymmärtävät, että heidän aineistoillaan olisi käyttöä muuallakin ja suhtautuvat siihen positiivisesti

Käytön mahdollistaminen vaatii kuitenkin monenlaisia hallinnointiin, lupamenettelyyn, tekniikkaan ja käytön tukeen liittyviä ratkaisuja, jotka maksavat: nämä menot olisi nähtävä kansallisen tietoinfrastruktuuri-investointina

3.4.Tutkimusrahoitusta myös aineistohankintoihin

Tutkijoiden kannalta aineistojen käyttö- tai hankintakustannuksen muodostuminen ei ole aina läpinäkyvää

Tutkimusrahoittajat eivät aina ole varautuneet siihen, että rahoitushakemuksiin sisältyy merkittäviä aineistokustannuksia (jotka vielä ovat "kaikki tai ei mitään" -tyyppiä)

3.5. Aineistojen käytön opastuksen parantaminen

Tätä työtä tekevät jo omilla tahoillaan

- Tilastokeskus
- FSD
- Rekisteritutkimuksen tukikeskus ReTKi

Lisäksi tulee lähiaikoina mietittäväksi, miten yhdistyvä Stakes/KTL organisoi omat tietovarantonsa ja miten järjestää niitä koskevan opastuksen.

4. Haaveet, pelot ja todellisuus aineistojen käytössä: yhteis-ymmärrystä ja sen puutetta

4.1. Tutkijapuolen haaveet, pelot ja todellisuus

4.2. Tilastoaineistojen kerääjän haaveet, pelot ja todellisuus

4.3. Mitä asioita eri osapuolten on toisistaan vaikea ymmärtää?

4.1. Tutkijan haaveet, pelot ja todellisuus

Tutkijahaave: vapaa pääsy aineistoihin niitä yhdistellen, ilman suuria kustannuksia

Tutkijapelko: "siilon" yksinoikeuden kutistuminen

Tutkijatodellisuus: sekava maailma seinineen ja porsaanreikineen, aineistokustannuksiin ei saa tutkimusrahoituslähteistä riittävää rahoitusta (rahoittajien puutteellinen ymmärrys), tietosuojamääräyksiä pidetään usein kohtuuttomina

4.2. Tilastoaineiston kerääjän haaveet, pelot ja todellisuus

Tilastoaineiston kerääjän haave: tietosuojan turvaava tekninen ratkaisu (esim. "Tanskan malli", tutkimuslaboratorio)

Tilastoaineiston kerääjän pelko: aineistojen kontrolloimaton käyttö, tietosuojaluottamuksen menetys (tietoja yhdistellen syntyy tunnistamisvaaroja), infra-kustannusten lankeaminen tilastonpitäjän kannettaviksi

Tilastoaineiston kerääjän todellisuus: työläästi hallinnoitava kirjavien käytäntöjen kenttä

4.3. Mitä asioita aineistojen avoimemman käytön osapuolten on vaikea ymmärtää toisistaan?

Tutkijapuolella:

- Tietosuojamääräysten tiukkuutta ja tilastoväen huolta tunnistamisvaarasta
- Aineistojen luovuttamiseen liittyviä kuluja
- Aineistopyyntöjen käsittelyn ja toimitusten hitautta

Tilastoaineistojen kerääjän puolella:

- Kyltymätöntä datanhimoa (paljon jää käyttämättä)
- Melko vähäistä tietosuojahuolta (erityisesti yhdistelyt)
- Haluttomuutta raportoida, mitä aineistoilla on tehty ja kuka on käyttänyt

5. Kaikesta tästä huolimatta: tietä riittää myös eteenpäin!

Usein jää huomaamatta, että monet tahot, jotka yksittäisissä asiakohdissa näyttävät olevan vastakkaisilla käsityskannoilla, kuitenkin tekevät vakavaa työtä aineistojen käytön helpottamiseksi ja avoimemman käytön edistämiseksi

=> TUSKIN OLLAAN MITENKÄÄN UMPIKUJASSA, mutta aikaa ja kärsivällisyyttä vaaditaan edelleen!