TAMPEREEN YLIOPISTOSTA VALMISTUNEIDEN SIJOITTUMISSEURANTA -AINEISTOSARJAN ANONYMISOINTISUUNNITELMA

Huom. esimerkki perustuu Tietoarkistoon arkistoituun aineistosarjaan: Tampereen yliopistosta valmistuneiden sijoittumisseuranta.

SISÄLTÖ:

Anfin tekijä(t): Tietoarkisto, Annika Sallinen

Anonymisoinnin toteuttaja(t): Tutkijat ja Tietoarkiston datan käsittelijät

1. AINEISTON TIEDOT ENNEN ANONYMISOINTIA

Tampereen yliopiston ura- ja rekrytointipalvelut on seurannut vastavalmistuneiden työllistymistä tutkimuksilla, joissa tarkastellaan Tampereen yliopistosta valmistuneiden työllisyystilannetta noin vuoden kuluttua valmistumisesta. Alla on esitetty anonymisoinnin ratkaisujen taustalla olevat tekijät, jotka perustuvat Tietoarkiston aineistonhallinnan käsikirjan ohjeisiin.

Perusjoukko: Vuosittain toteutetun aineistosarjan perusjoukkona on Tampereen yliopistossa ylemmän korkeakoulututkinnon tai lääketieteen lisensiaatin tutkinnon suorittaneet henkilöt, jotka ovat valmistuneet noin vuosi sitten kyselyn tekohetkestä. Kunkin vuoden perusjoukon tietoja kannattaa tarkastella Opetushallinnon tilastopalvelusta Vipusesta[1]. Vipusessa opiskelijamääriä voi tarkastella yliopiston, koulutusalan, sukupuolen, opintojen aloitusvuoden, kansalaisuuden ja iän mukaan. Anonymisoinnin näkökulmasta huomiota kannattaa kiinnittää erityisesti koulutusalojen suuruuteen ja sukupuolijakaumaan. Esimerkiksi vuonna 2016 ylemmän korkeakoulututkinnon suorittaneita oli lukumäärältään 1139. Koulutusaloittain valmistuneiden määrät vaihtelevat tietojenkäsittelyn ja tietoliikenteen (ICT) 42 valmistuneesta 462 yhteiskunnallisilta aloilta valmistuneeseen. Sukupuolijakauma on hyvin epätasainen esim. terveystieteistä valmistui vuonna 2016 vain 4 miestä.

Otanta: Kysely on kokonaisaineisto ja vastausprosentti on vaihdellut eri vuosina ollen kuitenkin aina yli 50 prosenttia.

[1] Vipunen on opetushallinnon tilastopalvelu. Vipusesta saa tilasto- ja indikaattoritietoa eri sektoreiden koulutuksesta ja koulutuksen jälkeisestä sijoittumisesta, korkeakouluissa tehdystä tutkimuksesta sekä väestön koulutusrakenteesta ja opiskelijoiden sosioekonomisesta taustasta. Ks. vipunen.fi

Aineistojen sisältö: Aineiston kysymykset käsittelevät työtilannetta ja työhistoriaa, työllistymiseen vaikuttaneita tekijöitä ja tyytyväisyyttä työtilanteeseen, tutkintoon sekä alumnitoimintaan. Taustamuuttujat ovat pääosin samat jokaisessa aineistossa.

Aineistoissa on havaittu seuraavia epäsuoria tunnisteita:

- koulutusala, koulutusohjelma, pääaine, laitos ja tutkinto
- opintojen aloitusvuosi
- nykyinen päätyönantaja
- ammatti-, tehtävä- tai virkanimike
- opintojen aloitusajankohta (syksy/kevät)
- edellinen tutkinto
- kansalaisuus (suomi/muu)
- lääni
- maakunta
- avomuuttujien tiedot

Aineisto sisältää myös avomuuttujia, joista suurin osa on strukturoitujen kysymyspattereiden jälkeen tulevia muu, mikä -avokysymyksiä esim. ”Nykyinen päätyönantajasi on: Muu mikä?”. Avomuuttujissa on satunnaisesti tietoa pääaineesta, opintojen aloitusvuosia, vaihto- ja harjoittelukohteiden tarkkoja nimiä, sairauksia, opettajien nimiä, tutkielmien aiheita ja muita satunnaisia epäsuoria tunnisteita. Anonymiteetin kannalta on hyvä, että sarjan taustatiedoissa ei ole ikämuuttujaa, sillä ääri-iät toimivat usein hyvinä tunnisteina.

Mitä aineiston tietoja yhdistelemällä henkilö saattaa olla tunnistettavissa? Yksittäistä henkilöä ei voi suoraan tunnistaa aineistosta saatujen tietojen perusteella, mikäli hän ei ole antanut avomuuttujissa suoria tunnisteita, kuten harvinaista ammatti-, tehtävä- tai virkanimikettään tai omaa nimeänsä avomuuttujissa.

Sisältääkö aineiso kolmansiin henkilöihin liittyviä tietoja ja voiko niiden perusteella tunnistaa henkilöitä? Kysymyksissä ei suoraan tiedustella kolmansiin henkilöihin liittyviä tietoja. Aineistot saattavat kuitenkin sisältää tietoja kolmansista henkilöistä avomuuttujissa esim. mainintoja yliopiston henkilökunnasta.

Ovatko aineiston tiedot sensitiivisiä? Aineiston kysymykset eivät suoraan kysy erityisiin tietoryhmiin kuuluvia tietoja, mutta avomuuttujissa saattaa olla satunnaisesti esim. terveyteen liittyviä tietoja kuten sairauksia. Kulttuurisessa kontekstissa sensitiivisiksi tiedoiksi voidaan arvioida työttömyyteen liittyvät tiedot.

Aineiston ikä: Aineistoja on Tietoarkistossa vuodesta 1998 lähtien ja uusimmat ovat vuodelta 2016. Vanhimpien aineistojen kohdalla tutkittavien tietojen voidaan olettaa muuttuneen. Kaikki aineistot anonymisoidaan silti samalla tarkkuudella, koska 90-luvulla valmistuneiden uratietoja tietoja voidaan olettaa löytyvän netistä esimerkiksi Linkedinistä.

Vastaajista muualta saatavat tiedot: Vastaajajoukkoon kuuluvista henkilöistä on tietoja esim. sosiaalisessa mediassa (esim. some-tilit, ryhmäjäsenyydet, LinkedIn), ja pro gradut ovat saatavilla usein yliopistojen nettisivuilta tekijä- ja koulutusalatietoineen. Lisäksi opiskelijamääristä löytyy tietoa Vipusesta, jossa määriä voi tarkastella yliopistoittain sekä iän, kansalaisuuden, koulutusalan ja sukupuolen mukaan. Mikäli urkkija epäilee henkilön opiskelevan yliopistossa hän voi tarkistaa asian yliopiston opintotoimistosta ja tieto on annettava, jos henkilö on antanut siihen luvan.

Ulkopuoliselle henkilölle yksittäisen henkilön tunnistaminen voi onnistua yhdistämällä aineiston tietoja ulkopuolelta saataviin tietoihin. Tämä on mahdollista esimerkiksi sosiaalisen median tietojen avulla, varsinkin jos henkilöllä on saatavilla esim. Facebookista tiedot opintojen pääaineesta, yliopistosta, opintojen aloitusvuodesta, sukupuolesta, nykyisistä ja entisistä työpaikoista, oppilaitoksista ja tieto kotipaikkakunnasta.

Käytettävyys vs. anonymiteetti: Aineiston käytettävyyden kannalta on tärkeämpää pyrkiä jättämään numeerisia muuttujia avomuuttujien sijaan. Eli avomuuttujia voi 'uhrata'. Numeerisista epäsuorista tunnisteista pyritään jättämään merkityksellisimmät, joita tässä tutkimuksessa ovat sukupuoli ja koulutusala. Tutkimuksellisesti voi olla mielenkiintoista jättää kansalaismuuttuja, jotta pystytään tutkimaan ulkomaalaisia opiskelijoita, mutta koska heidän määränsä on suhteellisen pieni koulutusaloittain, muuttuja voi sisältää tunnistamisriskin.

2. ANONYMISOINTIOHJEET PERUSTELUINEEN

Aineistoon jätetään seuraavat taustamuuttujat:

- sukupuoli
- koulutusala (luokiteltuna)
- tutkinto (luokiteltuna)
- tiedekunta/yksikkö (luokiteltuna)
- opintojen aloitusvuosi (luokiteltuna)
- lääni/suuraluejako NUTS 2
- edellinen tutkinto

1. **Poistetaan numeerisista muuttujista:**

    - pääaine
    - koulutusohjelma
    - laitos, kansalaisuus
    - koulutuskieli
    - suuntautumisvaihtoehto
    - maakunta

Perustelut: Valmistuneiden anonymiteetti säilyy parhaiten rajoittamalla tietoa pääaineesta ja tutkinto-ohjelmasta, kansalaisuudesta ja nykyisestä työpaikasta, sillä muihin tietoihin yhdistettynä ne antavat valmistuneesta liian yksilöiviä tietoja. Kansalaisuusmuuttuja poistetaan, koska Tampereen yliopistossa valmistuu vuosittain maisteriksi keskimäärin noin 60 ulkomaalaista opiskelijaa, ja eri aloilla niitä on n. 1–12 (katsottu Vipusen tilastoja vuosilta 2000–2016).

2. **Koulutusala, tutkinto ja yksikkö:** luokitellaan tarvittaessa. Varmista, että jokaiseen koulutusalaan, tutkintoon ja yksikköön jää tarpeeksi vastaajia. Sisään otettujen ja valmistuneiden opiskelijoiden määriä yliopistoissa koulutusaloittain ja pääaineittain voi tarkastella Opetushallinnon Vipusesta. Tunnisteellisuutta vähennetään luokittelemalla pienimmät koulutusala-, tutkinto- ja yksikköarvot. Tällaisia ovat esimerkiksi

    a. tutkintomuuttujan teatteritaiteen maisterin tutkinto, joka yhdistetään filosofian maisterin tutkintoon
    b. psykologia, joka yhdistetään yhteiskuntatieteelliseen alaan ja
    c. terveystiede, joka yhdistetään lääketieteeseen.

    Luokittelussa noudatetaan Tilastokeskuksen koulutusalaluokitusta. Puolestaan yksikkömuuttujassa terveystieteiden yksikkö tulee luokitella yhteen lääketieteen yksikön kanssa.

    Jos kyseisenä valmistumisvuonna sukupuolijakauma on ollut jollain alalla epätasainen <5, muuttujia tai havaintojen arvoja tulee muokata. Arvot voi poistaa, jos sysmis-arvoja on tarpeeksi jo valmiina. Jos edellinen ei ole mahdollista, toinen vaihto anonymisoinnille on aineiston sekoittaminen, eli enemmistön sukupuolen edustajia muutetaan vastakkaiseksi sukupuoleksi tai vähemmistön muuttaminen enemmistön sukupuoleen. Esimerkiksi jos vuonna 2010 on valmistunut Vipusen mukaan terveystieteestä kolme miestä, joista kaksi on vastannut kyselyyn, arvot voidaan muuttaa naisiksi, jos naiseksi valmistuneita on suhteellisessa enemmän. Käytettäessä satunnaistavia menetelmiä vaikutukset tilastollisiin analyyseihin tulee jäädä pieniksi. Jatkokäyttäjille on hyvä mainita, että aineistossa kaksi miestä on muutettu naiseksi miesten vähäisen määrän vuoksi.

    Perustelut: Koulutusala- ja tutkintotiedoista luokitellaan yhteen pienten sisäänottojen alat. Esimerkiksi teatterityön maisterit on yhdistetty humanististen maistereiden kanssa samaan luokkaan, koska teatterityön on katsottu olevan sellainen pienen opiskelijamäärän sisäänoton ala, jolla tunnisteellisuus on esim. näyttelijöiden julkisuuden vuoksi suurempi kuin muilla aloilla.

    Anonymiteetin kannalta tärkeää on huomioida sukupuolten väliset epätasaiset jakaumat eri aloilla. Tämä pyritään huomioimaan aloitusvuoden luokittelulla ja mahdollisilla yksikkö-, tutkinto ja koulutusalaluokitteluilla sekä avovastausten anonymisoinnilla. Mikäli sukupuolijakaumat edelleen jäävät pieniksi, tietoja merkitään puuttuviksi tai sekoitetaan sotkevilla menetelmillä. Arvojen poistaminen tai sotkevien menetelmien käyttö katsotaan järkevämmäksi jatkokäytön kannalta verrattuna koulutusalamuuttujan luokitteluun edelleen karkeammaksi tai muuttujan kokonaan poistamiseen. Koulutusalan arvojen edelleen yhdistäminen tuottaisi tutkimuksellisesti käyttökelvottomia luokkia.

3. **Opintojen aloitusvuosi:** luokitellaan kahden vuoden välein ja muodostetaan kaikista pisimmin opiskelleista oma vähintään 20 opiskelijan ryhmä esim. ”ennen vuotta xxxx aloittaneet”.

Luokittelu on tehty yleisimmin seuraavasti: kaksi ensimmäistä luokkaa kahden vuoden välein, kolmas kolmen vuoden välein ja neljänteen loput. Luokittelu aloitetaan tuoreimmasta vuodesta. Vanhemmissa aineistoissa muuttuja on luokiteltu kolmen vuoden välein arvojen suuremman vaihtelun vuoksi.

Perustelut: Opintojen aloitusvuosi luokitellaan, jotta ei voida paikantaa suoraan, minä vuonna opiskelijat ovat aloittaneet opiskelut. Opintojen aloitusvuosi on tärkeä tieto opiskelijoille verrattuna valmistumisvuoteen, sillä usein ihmiset muistavat samana vuonna opintonsa aloittaneet henkilöt. Joillain aloilla sisään-ottomäärät ovat myös sen verran pieniä ja sukupuoleltaan epätasaisesti jakautuneita, että vuosien yhdistäminen estää harvinaisten henkilöiden tunnistamisen.

4. **Asuinmaakuntamuuttuja:** luokitellaan NUTS2-suuraluejaon mukaisesti: 1 Helsinki-Uusimaa; 2 Etelä-Suomi; 3 Länsi-Suomi; 4 Pohjois- ja Itä-Suomi ja 5 Ahvenanmaa-Åland. Vanhemmissa sarjan aineistossa maakunnat luokitellaan lääneittäin: 1 Etelä-Suomen lääni; 2 Länsi-Suomen lääni ja 3 Oulun, Lapin ja Itä-Suomen läänit. Läänit lakkautettiin käytöstä 2011.

   Perustelut: Aineistoon ei voi jättää asuinpaikan aluemuuttujaksi maakuntaa, koska tieto saattaisi muodostua tunnistamisen mahdollistavaksi tekijäksi, jos urkkija epäilee tunnistavansa henkilön. Maakunnan voi usein päätellä netistä saatavista työpaikkatiedoista. Tietyt maakunnat voivat olla myös harvinaisia muuttokohteita valmistumisen jälkeen esim. harvaan asutummat tai kaukaisemmat seudut.

5. **Avomuuttujista poistetaan** kaikki muut avomuuttujat pl. tyytyväisyyttä tai tyytymättömyyttä Tampereen yliopiston opetukseen käsittelevät muuttujat. Em. muuttujat jätetään sillä, niistä ei ole luokiteltua tietoa ja anonymisointi on toteutettavissa kohtuullisin resurssein. Aineistosta poistetaan siis esim. ammattinimike, alumnitoimintaan ja harjoitteluun liittyvät avokysymykset ja sellaiset avomuuttujat, joista on jo luokiteltua tietoa ja sisältävät anonymisointia vaativia tarkkoja tietoja vastaajan työstä tai elämäntilanteesta. Tällaisia ovat mm. muu, mikä -avokysymykset: Muuten työelämän ulkopuolella, miten? Nykyinen päätyönantaja, muu mikä? Mikä seuraavista lähinnä kuvaa päätyötäsi: mikä muu? Työllistymisvaikeudet, muu mikä? ja Kielimuuttuja (mitä kieliä osaa?).

   Perustelut: Aineistosta poistetaan avomuuttujat, jotka voivat antaa liian yksityiskohtaisia tai harvinaisia tietoja nykyisestä elämäntilanteesta ja työnkuvasta. Esimerkiksi ammattinimike poistetaan, sillä osa ammattinimikkeistä voi olla käytössä Suomessa vain muutamalla henkilöllä. Poisto voidaan tehdä, koska aineistoon jää edelleen tietoa ammatista ja työpaikasta työnkuva- ja työnantajatyyppimuuttujiin. Jos tutkija haluaa tutkia ammattinimikkeitä, tiedot löytyvät Tampereen yliopiston uraseuranta-palvelun nettisivuilta. Nettisivuilta saatavista tiedoista ei voi tunnistaa yksittäistä opiskelijaa, sillä ne on koottu useamman vuoden opiskelijoiden tiedoista, ei yksittäisten vuosien perusteella.

   Aineistosta poistetaan myös alumnitoimintaan liittyvät avomuuttujat, sillä tietojen katsotaan hyödyntävän eniten yliopistoa, eivät jatkokäyttäjiä. Harjoitteluun liittyvä avomuuttuja poistetaan, sillä siitä on jo ei-tunnisteellista luokiteltua tietoa muissa muuttujissa. Avomuuttujien poistolla on haluttu vähentää myös niiden yksityiskohtaisesta läpikäynnistä ja muokkauksesta syntyvää työn määrää.

6. **Jätetyt avomuuttujat** anonymisoidaan niin, että niistä ei voida yksilöidä henkilöä tai kolmansia henkilöitä harvinaisten tapahtumien, henkilön ominaisuuksien, titteleiden, pääaineen tms. kautta. Avovastaukset eivät saa paljastaa tarkkaa pääainetta pienimmiltä aloilta, opintojen aloitusvuotta, erillisiä yliopisto-opetuksen paikkakuntia (Pori, Seinäjoki), harvinaisia maisteriohjelmia, graduohjaajien nimiä tai ammattinimikkeitä. Pienten sisäänoton aloja ovat esimerkiksi logopedia, näyttelijän opinnot ja filosofia. Yliopiston henkilökunnan nimet voidaan jättää, jos henkilöstä puhutaan positiiviseen sävyyn. Herjaavista tai loukkaavista kommenteista nimi poistetaan.

   Esimerkkejä avomuuttujien anonymisoinneista (esimerkit ovat keksittyjä):

   Gradun ohjaajani (Jasper Jensson) ohjasi ryhmäänsä hyvin → Gradun ohjaajani ([nimi poistettu]), ohjasi ryhmäänsä hyvin

   Opiskeluaikana sairastuin vakavasti ja sen vuoksi opintoni kestivät 7v. → Opiskeluaikana [syy poistettu] ja sen vuoksi opintoni kestivät [tavoitekestoa enemmän].

   Keskeytin opintoni vuonna 2003, vuodet 2004-2005 olin vaihdossa Kiinassa ja harjoittelussa ja valmistuin lopulta 2015 → Keskeytin opintoni vuonna [x], vuodet [2 vuoden aika] olin vaihdossa [kohteessa x] ja harjoittelussa ja valmistuin lopulta 2015

3. PALJASTUMISRISKIN ARVIOINTI ANONYMISOINNIN JÄLKEEN

Voiko anonymisoinnin jälkeen aineiston tietoja yhdistää ulkopuolisiin tietoihin? Anonymisoinnin jälkeen aineiston tietojen yhdistäminen toisiin aineistoihin kuten muihin Tampereen yliopiston opiskelijoilleen teettämiin kyselyihin, pro gradu -tutkielmiin tai muualla oleviin tietoihin on mahdollista, mutta tunnistaminen niiden perusteella erittäin epätodennäköistä ja täsmällisen osuman löytäminen olisi hyvin vaikeaa. Aineiston havaintoja ei voi yhdistää minkään koodiavaimen avulla toisiinsa.

Henkilöiden harvinaiset ominaisuudet tms., tarkka työnkuva tai työpaikan muut tiedot ovat anonymisoitu aineistoista, jotta yksittäisen henkilön tietoja ei voi etsiä netistä esim. Linkedinistä tai Facebookista. Opinnäytetöiden perusteella tunnistaminen on jokseenkin mahdotonta, sillä kyselyssä ei kysytä tietoja gradusta, ja graduntekovuosi ei ole aina sama kuin henkilön valmistumisvuosi. Pro gradut saattavat antaa osviittaa henkilöistä vain sellaisissa tapauksissa, joissa koulutusalan sukupuolijakauma on hyvin epätasainen, mutta anonymisoinnissa otettiin sukupuolijakaumat huomioon.