Viite

Viite

Asia

Tilastokeskuksen yleiset toimintaperiaatteet koskien verkkoharavointia

Tilastokeskuksen haravoidessa internetistä tietoja tilastotuotantoa varten noudatetaan seuraavia yleisiä periaatteita:

Verkkoharavoinnilla hankittavan aineiston käytöstä tiedonkeruupäätös ja aineisto kuvataan Tiedonkeruurekisteriin. Tiedonantovelvollisuuteen perustuvan verkkoharavoinnin yhteydessä noudatetaan tilastoviranomaisen normaaleja tiedonanto- ja neuvotteluvelvoitteita. Verkkoharavoinnin keinoin hankitun aineiston laatuun liittyvät riskit tunnistetaan ja arvioidaan ennen verkkoharavointia koskevien päätösten tekemistä. Verkkoharavoinnin käyttäminen tiedonhankintatapana ilmoitetaan tilaston julkaisemisen yhteydessä.

Lainmukaisuus. Lainsäädäntö huomioidaan suunnitteluvaiheesta alkaen, ja sitä noudatetaan täysimääräisesti. Mahdollisia muutoksia oikeustilassa seurataan.

Läpinäkyvyys. Verkkoharavoinneista ilmoitetaan julkisesti Tilastokeskuksen kotisivuilla, ja annetaan samalla seuraavat tiedot: haravoinnin tarkoitus, haravoinnin kohteena olevat tietotyypit sekä yhteystiedot, joihin verkkosivuston ylläpitäjä voi ottaa yhteyttä. Tiedonantovelvollisuuteen perustuvassa tiedon keräämisessä huomioidaan tilastolain mukainen tiedottamisvelvollisuus. Jos aineisto sisältää henkilötietoja, henkilötietojen käsittelyä koskevat tiedot julkaistaan avoimesti ja helposti saatavasti Tilastokeskuksen kotisivuilla.

Vähimmän haitan periaate. Verkkoharavointi toteutetaan kaikin puolin siten, että haravoinnista aiheutuu verkkosivuston toiminnalle ja sen omistajille mahdollisimman vähän haittaa ja kustannuksia.

Oikeus kieltää. Verkkosivuston ylläpitäjille annetaan oikeus kieltää haravointi ottamalla yhteyttä. Kieltopyyntöjä kunnioitetaan, ja ne listataan.

Tilastointiperiaatteiden noudattaminen. Tilastojen laatimiseen sovellettavia menettelytapoja ja periaatteita sekä tilasto- ja tutkimusalan ammattieettisiä periaatteita noudatetaan.

Käyttöehtojen tarkistaminen. Verkkoharavointi kohdistetaan ainoastaan sellaisille verkkosivustoille, joiden käyttöehdoissa verkkoharavointia ei ole nimenomaisesti kielletty tai kielto on selkeästi rajattu koskemaan kaupallista toimintaa. Epäselvissä tilanteissa voidaan ottaa yhteyttä sivuston ylläpitäjään. Jos vastausta ei saada kohtuullisen ajan kuluessa, sivustoa voidaan haravoida.

Yleisten periaatteiden lisäksi Tilastokeskus noudattaa verkkoharavointia toteuttaessaan seuraavia käytännön toimintaperiaatteita:

Tietojen tarpeellisuus. Haravoitavien tietojen on oltava perustellusti tarpeellisia tilastojen laatimiseen ja tuotettava lisäarvoa tilastotuotantoon.

Käyttötarkoitus. Verkkoharavoinnin avulla kerättyjä aineistoja voidaan luovuttaa ainoastaan Tilastolain 13 §:n mukaisiin tarkoituksiin.

Identiteetin paljastaminen (user agent string). Verkkosivustolle ilmoitetaan Tilastokeskuksen identiteetti, yhteyspiste yhteydenottoa varten sekä linkki verkkoharavointia koskevaan ilmoitukseen Tilastokeskuksen kotisivuille.

Kuormittamisen minimointi. Verkkosivustoja ei kuormiteta liiallisilla kyselyillä, vaan ne tehdään kohtuullisin intervallein. Haravointi ajoitetaan mahdollisuuksien mukaan sellaisiin ajankohtiin, että sivustolla ei ole odotettavissa runsaasti liikennettä. Ylimääräisiä kyselyitä ei tehdä.

Etukäteiskuuleminen poikkeustapauksissa. Verkkosivuston ylläpitäjää kuullaan tapauksissa, joissa verkkoharavointi olisi poikkeuksellisen laajamittaista tai kuormittavaa.

Tilannekohtainen harkinta. Verkkoharavoinnin tilannekohtainen tarkoituksenmukaisuus selvitetään ennen haravoinnin aloittamista.

Robots.txt. Jos verkkosivustolla on verkkoharavoinnin kieltävä robots.txt-tiedosto, sitä kunnioitetaan. Robots.txt-tiedostosta poikkeamiseen tarvitaan etukäteen kirjallinen lupa verkkosivuston ylläpitäjältä.

Jos verkkoharavoinnin keinoin hankittua aineistoa hankitaan kolmannelta osapuolelta, noudatetaan seuraavia periaatteita:

Tilastokeskus voi hankkia verkkoharavoitua aineistoa ainoastaan siinä tapauksessa, että aineiston tuottaneen toimittajan toimintatavat eivät ole ristiriidassa Tilastokeskuksen verkkoharavointia koskevien toimintaperiaatteiden kanssa. Ennen sopimuksen laatimista on varmistettava, että hankinta ei johda välillisestikään tässä ohjeessa määriteltyjen toimintaperiaatteiden rikkomiseen tai luvattomasti haravoitujen tietojen hankkimiseen.

Toimittajan tulee osoittaa, että aineisto on luvallisesti hankittua joko siten, että sitä ei ole haravoitu sivustojen käyttöehtojen vastaisesti, tai että haravoinnista ja aineiston myyntioikeudesta on sovittu haravoinnin kieltävien verkkosivustojen ylläpitäjien kanssa. Lisäksi aineiston on oltava muutoinkin eettisesti kestävällä tavalla hankittua, eikä se saa olla esimerkiksi tekijänoikeuden suojaamasta tietokannasta kopioitua.

Henkilötietoja sisältävää aineistoa ei lähtökohtaisesti voida hankkia valmiina, vaan haravoinnin tulisi tapahtua Tilastokeskuksen nimenomaisesta toimeksiannosta siten, että henkilötietojen käsittelystä on sovittu Tilastokeskuksen ja toimeksisaajan välillä.

Marjo Bruun
Pääjohtaja

Timo Koskimäki
Tilastotuotannon ylijohtaja

Liite: Verkkoharavoinnin toimintaperiaatteet Tilastokeskuksessa

Verkkoharavoinnin toimintaperiaatteet Tilastokeskuksessa

1. Johdanto

Digitalisoituminen ja tietomäärien kasvu haastaa ja mahdollistaa myös tilastotuotantoa- ja sen kehittämistä uudella tavalla. Internetissä on saatavilla yhä laajemmin niin yritysten kuin henkilöidenkin tietoja, joita voitaisiin hyödyntää myös tilastotuotannossa.

Vaikka internetistä poimittujen tietojen käyttö haastaa niin tilastotuotantojärjestelmiä kuin käytettäviä menetelmiä, niiden käytölle on myös paljon hyviä perusteita.

Tilastolain mukaan tilastojen laadintaan tarvittavat tiedot tulisi kerätä mahdollisimman tehokkaasti ja tiedonantajien rasite minimoiden: mikä olisikaan tähän parempi keino, kuin käyttää internetistä jo valmiiksi löytyviä tietoja. Verkosta haravoidut tiedot voisivat olla ratkaisu myös tiedonkeruiden laskeviin vastausasteisiin, ja niiden avulla voitaisiin vähentää työvoimaintensiivisiä ja näin ollen kalliita suoria tiedonkeruita. Myös tilastotuotantoa voitaisiin näin tehostaa. Esimerkiksi kuluttajahintaindeksissä on hyvin tuloksin kokeiltu hintatietojen haravointia yritysten verkkosivuilta.

Uusia tietotarpeita, joiden täyttämiseen pohditaan aineistojen hankintaa, syntyy koko ajan. Valmiita rekistereitä ei aina löydy, tai tietojen kerääminen suorilla tiedonkeruilla olisi liian kallista tai työlästä toteuttaa. Verkkoharavointi voisi tarjota ratkaisuja myös tällaisiin tilanteisiin.

Vaikka moni seikka puoltaa internetistä haravoitujen tietojen hyödyntämistä, ei tietojen käyttäminen tilastoinnissa ole ongelmatonta. Verkkoharavoinnissa törmätään tiedon laatua koskevien puutteiden lisäksi sekä eettisiin että juridisiin ongelmiin. Voiko tilastotuotannossa käyttää verkkosivuilta löytyviä tietoja, joiden haravointi on verkkosivuston käyttöehtojen mukaan kielletty? Entä tuleeko tiedon kohteilta pyytää lupa heidän tuottamansa aineistojen käyttöön? Kuinka suhtautua tilanteeseen, jossa verkkosivuston ylläpitäjä sallii verkkoharavoinnin ainoastaan maksua vastaan?

Samoja haasteita pohditaan eurooppalaisessa tilastoyhteistyössä. Ensimmäinen verkosta haravoitujen tietojen käyttöä koskeva linjaus (ESS Web scraping policy template) julkaistiin heinäkuussa 2019, ja se on yhtenäinen tässä dokumentissa esitettyjen käytäntöjen kanssa.

Tähän ohjeeseen on koottu Tilastokeskuksen yleiset toimintaperiaatteet koskien internetistä haravoitujen tietojen käyttöä tilastotuotannosta. Dokumenttia päivitetään, jos lainsäädäntö muuttuu tai verkkoharavointia koskevaa uutta tietoa tai ohjeita tulee muutoin saataville.

2. Verkkoharavointiin vaikuttava lainsäädäntö

Tilastojen laadintaa ohjaa tilastolaki (280/2004). Tilastolakiin tai muuhun lainsäädäntöön ei sisälly varsinaisia verkkoharavointia koskevia säännöksiä. Ainoa poikkeus on laki kulttuuriaineistojen tallettamisesta ja säilyttämisestä (1433/2007), jossa Kansalliskirjaston tehtäväksi on säädetty hakea ja tallentaa yleisön saatavilla olevaa verkkoaineistoa. Verkkoharavointia tilastotuotannossa koskevia vakiintuneita tulkintoja tai oikeuskäytäntöä ei ole, ja käytännöt ovat kansainväliselläkin tasolla vasta muotoutumassa. Verkkoharavointia onkin tällä hetkellä tarkasteltava yleislainsäädännön ja tilastojen laadintaa koskevan lainsäädännön perusteella. Verkkoharavoinnin lainmukaisuutta arvioitaessa on erityisesti huomioitava kolme näkökulmaa: tekijänoikeudet, tietosuoja ja käyttöehdot.

Verkkosivustojen sisältöön mahdollisesti kohdistuvat **tekijänoikeudet** eivät sinänsä rajoita verkkoharavointia, sillä Tilastokeskuksen mielenkiinnonkohteena on verkkosivustoilla julkaistujen teosten, tietojen tai aineistojen sijasta niistä välittyvä informaatio. Tekijänoikeudet eivät vaikuta esimerkiksi yritystä koskevien tietojen haravointiin yrityksen kotisivuilta. Jotkin verkkoalustat katsovat kuitenkin muodostavansa tietokannan. Tekijänoikeuslain (404/1961) mukaan huomattavaa panostusta edellyttäneen tietokannan valmistajalla on EU:n tietokantadirektiivin mukaisesti yksinoikeus tietokantaan sekä oikeus estää tietokannan kopiointi tai uudelleenkäyttö (sui generis -oikeus). Sui generis -suojan saaminen on kuitenkin Euroopan unionin tuomioistuimen oikeuskäytännössä edellyttänyt tietokannan valmistamiselta niin merkittäviä rahallisia tai ajallisia voimavaroja, että äärimmäisen harva verkkoalusta täyttää nämä kriteerit. Tällöinkin on huomioitava, että jos tietokanta on yleisölle avoin, tietokannan laatija ei tuomioistuimen mukaan voi kieltää ulkopuolisia etsimästä sieltä tietoja.

Jos haravoitavat sivut sisältävät henkilötietoja, on ennen verkkoharavoinnin aloittamista otettava huomioon myös **tietosuojalainsäädäntö**. Henkilötietojen käsittelystä säädetään EU:n yleisessä tietosuoja-asetuksessa (Euroopan parlamentin ja neuvoston asetus (EU) 2016/679) ja sitä täydentävässä tietosuojalaissa (1050/2018). Henkilöä, jota henkilötiedot koskevat, kutsutaan rekisteröidyksi. Erityistä huomiota tulee kiinnittää siihen, että haravoitavaan aineistoon sisältyvät henkilötiedot ovat tilaston tuottamisen näkökulmasta kaikilta osiltaan asianmukaisia, olennaisia ja tarpeellisia. Verkkosivuston lataaminen haravoinnin yhteydessä voi kuitenkin johtaa siihen, että poimintaan tarttuu muutakin kuin tilastotuotannossa tarvittavaa aineistoa. Jos tällainen tilapäinen, teknisen prosessin sivutuotteena tallentunut aineisto sisältää tahattomasti henkilötietoja, se on hävitettävä tai anonymisoitava viipymättä.

Tilaston tuottamiseen nähden ylimääräisten henkilötietojen haravointi on minimoitava, ja haravat on lähtökohtaisesti suunniteltava siten, ettei ylimääräisiä henkilötietoja tallenneta.

Rekisteröityjä ei tarvitse informoida henkilökohtaisesti heidän tietoihinsa kohdistuvasta verkkoharavoinnista, sillä se vaatisi kohtuutonta vaivaa ja estäisi tilastollisen tavoitteen saavuttamisen tietosuoja-asetuksen salliman poikkeuksen mukaisesti. Kollektiivinen informointi on tällaisessa tapauksessa riittävää, ja se tehdään ilmoittamalla verkkoharavoinnista ja siihen liittyvästä henkilötietojen käsittelystä Tilastokeskuksen kotisivuilla. Tietojen on katettava tietosuoja-

asetuksen vaatimukset, ja niiden on oltava selkeästi ilmaistuja ja helposti löydettävissä.

Jos haravoitava aineisto sisältää henkilötietoja, henkilötietojen käsittelyä koskevat tiedot julkaistaan avoimesti ja helposti saatavasti Tilastokeskuksen kotisivuilla.

Myös rekisteröityjen oikaisu-, tarkastus-, rajoittamis- ja vastustamisoikeutta voidaan tietosuojalain nojalla poikkeuksellisesti rajoittaa, jos tilaston tuottaminen tai tiedontarve sitä vaatii. Rekisteröidyn oikeuksista poikkeaminen edellyttää aina Tilastokeskuksen tietosuojavastaavan kuulemista ja vaikutustenarvioinnin toteuttamista.

Verkkosivustolla olevien tietojen käyttöä voidaan rajoittaa **käyttöehdoilla**. Jotta käyttöehdot tulisivat sovellettavaksi, on verkkosivuston käyttäjän hyväksyttävä ne esimerkiksi rekisteröitymisen tai sivustolla vierailun yhteydessä. Osa verkkoalustoista kieltää verkkoharavoinnin käyttöehdoissaan toteamalla esimerkiksi, että ”käyttäjällä ei oikeutta käyttää automaattisia järjestelmiä kappaleiden valmistamiseksi” tai että ”automatisoitujen ohjelmistojen käyttäminen tietojen kopiointiin internetsivustolta on kiellettyä”. Koska tällaisten ehtojen pääsääntöisenä rationa on ollut rajoittaa tietojen kaupallista hyödyntämistä, niiden merkityksestä tilastoviranomaisten harjoittamaan verkkoharavointiin on esitetty vaihtelevia tulkintoja. Vahva pääsääntö on kuitenkin se, että verkkosivustojen käyttöehdoissa ei ole otettu kantaa verkkoharavointiin.

3. Verkkoharavoinnin yleiset periaatteet

Verkkoharavoinnilla hankittavan aineiston käytöstä tehdään aina tiedonkeruupäätös ja aineisto kuvataan muiden hankittavien aineistojen kaltaisesti Tiedonkeruurekisteriin. Tiedonantovelvollisuuteen perustuvan verkkoharavoinnin yhteydessä noudatetaan tilastoviranomaisen normaaleja tiedonanto- ja neuvotteluvelvoitteita.

Verkkoharavoinnin keinoin hankitun aineiston laatuun liittyvät riskit tunnistetaan ja arvioidaan ennen verkkoharavointia koskevien päätösten tekemistä. Aineiston paikkansapitävyyttä ja ajantasaisuutta ei usein pystytä kaikilta osin tarkistamaan. Verkkoharavoinnin käyttäminen tiedonhankintatapana on tämän vuoksi ilmoitettava nimenomaisesti haravoidun aineiston pohjalta muodostetun tilaston julkaisemisen yhteydessä. Jos haravoituja tietoja halutaan käyttää esimerkiksi koneoppimisessa, on myös aineiston puolueettomuus pystyttävä varmistamaan.

Eettisen kestävyyden varmistamiseksi kaikessa Tilastokeskuksen harjoittamassa verkkoharavoinnissa on noudatettava seuraavia periaatteita:

Lainmukaisuus. Lainsäädäntö, tietosuoja mukaan lukien, huomioidaan jo haravoinnin suunnitteluvaiheessa, ja sitä noudatetaan täysimääräisesti. Mahdollisia muutoksia oikeustilassa (lainsäädäntö, oikeuskäytäntö, vakiintuneet tulkinnat) seurataan.

Läpinäkyvyys. Verkkoharavoinneista ilmoitetaan julkisesti Tilastokeskuksen kotisivuilla. Samassa yhteydessä annetaan kustakin verkkoharavoinnista seuraavat tiedot: haravoinnin tarkoitus, haravoinnin kohteena olevat tietotyypit sekä yhteystiedot, joihin verkkosivuston ylläpitäjä voi ottaa yhteyttä lisätietojen pyytämiseksi tai haravoinnin rajoittamiseksi. Jos kyse on tiedonantovelvollisuuteen perustuvasta tiedon keräämisestä, tietoja kerättäessä

huomioidaan lisäksi tilastolain mukainen tiedottamisvelvollisuus. Henkilötietojen käsittelyä koskevaa informointivelvollisuutta noudatetaan.

Vähimmän haitan periaate. Verkkoharavointi toteutetaan kaikin puolin siten, että haravoinnista aiheutuu verkkosivuston toiminnalle ja sen omistajille mahdollisimman vähän haittaa ja kustannuksia.

Oikeus kieltää. Verkkosivuston ylläpitäjille annetaan oikeus kieltää haravointi (opt-out) ottamalla yhteyttä Tilastokeskukseen. Kieltopyyntöjä kunnioitetaan, ja ne ilmoitetaan yhteiseen listaan (ns. black list).

Tilastointiperiaatteiden noudattaminen. Tilastojen laatimiseen sovellettavia tiedonkeruuta sekä tilastojen suunnittelua ja laadintaa koskevia menettelytapoja ja periaatteita noudatetaan myös verkkoharavoinnissa. Sama koskee Tilastokeskuksen toiminnan perustana olevia tilasto- ja tutkimusalan ammattieettisiä periaatteita.

Käyttöehtojen tarkistaminen. Verkkoharavointi kohdistetaan toistaiseksi ainoastaan sellaisille verkkosivustoille, joiden käyttöehdot on tarkistettu. Haravoinnin katsotaan olevan sallittua, jos sitä ei ole nimenomaisesti kielletty tai kielto on selkeästi rajattu koskemaan pelkästään kaupallista toimintaa. Epäselvissä tilanteissa voidaan ottaa yhteyttä sivuston ylläpitäjään. Jos vastausta ei saada kohtuullisen ajan kuluessa, sivustoa voidaan haravoida.

4. Verkkoharavoinnin käytännöt

Yleisten periaatteiden lisäksi Tilastokeskus noudattaa verkkoharavointia toteuttaessaan seuraavia käytännön toimintaperiaatteita:

Tietojen tarpeellisuus. Verkkoharavointi kohdistetaan vain sellaisiin tietoihin, jotka ovat perustellusti tarpeellisia tilastojen laatimisen kannalta. Tietojen on tuotava lisäarvoa tilastotuotantoon.

Käyttötarkoitus. Verkkoharavoinnin avulla kerättyjä aineistoja voidaan luovuttaa ainoastaan Tilastolain 13 §:n mukaisiin tarkoituksiin.

Identiteetin paljastaminen (user agent string). Verkkosivustolle ilmoitetaan Tilastokeskuksen identiteetti, yhteyspiste yhteydenottoa varten sekä linkki verkkoharavointia koskevaan ilmoitukseen Tilastokeskuksen kotisivuille.

Kuormittamisen minimointi. Verkkosivustoja ei kuormiteta liiallisilla kyselyillä, vaan ne tehdään kohtuullisin intervallein. Haravointi ajoitetaan mahdollisuuksien mukaan sellaisiin ajankohtiin, että sivustolla ei ole odotettavissa runsaasti liikennettä (esim. yöaikaan). Ylimääräisiä kyselyitä ei tehdä, vaan haravat toteutetaan siten, että ne hakevat ainoastaan tarpeellisia tietoja.

Etukäteiskuuleminen poikkeustapauksissa. Verkkosivuston ylläpitäjää kuullaan etukäteen tapauksissa, joissa verkkoharavointi olisi poikkeuksellisen laajamittaista tai kuormittavaa.

Tilannekohtainen harkinta. Verkkoharavoinnin tilannekohtainen tarkoituksenmukaisuus selvitetään ennen haravoinnin aloittamista. Tiedot voidaan hakea myös APIn kautta, jos sellainen on tarjottu.

Robots.txt. Jos verkkosivustolla on verkkoharavoinnin kieltävä robots.txt-tiedosto, sitä kunnioitetaan. Robots.txt-tiedostosta poikkeamiseen voidaan tarvittaessa kysyä lupaa verkkosivuston ylläpitäjältä (kirjallisesti). Haravointia ei tule aloittaa ennen myöntävän vastauksen saamista.

5. Kolmannen osapuolen haravoiman aineiston hankinta

Markkinoilla on yrityksiä, joiden liiketoiminta perustuu verkkoharavoidun aineiston myyntiin ja hyödyntämiseen. Mielenkiinnon kohteena olevien tietojen hankinta suoraan verkkoharavointia harjoittavilta kaupallisilta toimijoilta voi olla houkutteleva vaihtoehto. Kustannustehokkuuden selvittämisen lisäksi on kuitenkin varmistettava, että verkkoharavoinnissa noudatetaan samoja periaatteita riippumatta siitä, kuka – Tilastokeskus vai kolmas toimija – aineiston on hankkinut.

Ennen sopimuksen laatimista on varmistettava, että hankinta ei johda välillisestikään tässä ohjeessa määriteltyjen toimintaperiaatteiden rikkomiseen. Erityistä huomiota tulee kiinnittää siihen, että aineisto ei sisällä luvattomasti haravoituja tietoja. Toimittajan tulee joko osoittaa, että aineisto ei sisällä käyttöehtojen vastaisesti haravoituja tietoja, tai että haravoinnista ja aineiston myyntioikeudesta on sovittu haravoinnin kieltävien verkkosivustojen ylläpitäjien kanssa. Lisäksi aineiston on oltava muutoinkin eettisesti kestävällä tavalla hankittua (kuormittamisen minimointi, robots.txt, läpinäkyvä identiteetti), eikä se saa olla esimerkiksi tekijänoikeuden suojaamasta tietokannasta kopioitua.

Henkilötietoja sisältävää aineistoa ei lähtökohtaisesti voida hankkia valmiina, vaan haravoinnin tulisi tapahtua Tilastokeskuksen nimenomaisesta toimeksiannosta siten, että henkilötietojen käsittelystä on sovittu Tilastokeskuksen ja toimeksisaajan välillä.