Anonymisointi – askel kohti vapaammin hyödynnettävää tietoa

Tieto on valtaa, totesi englantilainen filosofi Francis Bacon jo noin 400 vuotta sitten. Tämä lausahdus ei voisi olla todempi digitaalisen tiedon kanssa. Esimerkkinä mainittakoon Venäjän oletettu vaikuttaminen USA:n presidentinvaaleihin. Erinäiset lait ja asetukset kuitenkin hankaloittavat tiedon vapaata saantia. Anonymisointi onkin ensimmäisiä askeleita kohti vapaammin hyödynnettävää tietoa.

TEKSTI: ANSSI JÄÄSKELÄINEN

Näistä lähtökohdista esittelen anonymisointia sekä kyllä- että ei-näkökulmista ja kerron lyhyesti ratkaisusta, joka Kaakkois-Suomen ammattikorkeakoulun (Xamkin) Digitaalisen tiedonhallinnan tutkimus- ja kehittämiskeskus Digitaliassa on toteutettu.

Anonymisoinnin perimmäisenä tarkoituksena on ihmisen yksityisyydensuojan turvaaminen. Tarkempi määritelmä riippuu lähteestä. Esimerkiksi tietosuoja.fi -palvelussa anonymisointi on määritelty seuraavasti: "Anonymisointi tarkoittaa henkilötietojen käsittelyä niin, että henkilöä ei enää voida tunnistaa niistä" ja tietosuojatyokalu.fi -palvelussa puolestaan näin: "Henkilötiedon tunnistettavuuden poistaminen siten, ettei se enää ole yhdistettävissä rekisteröityyn henkilöön edes lisätietojen avulla". Teoriassa helposta kuulostava asia on kaikkea muuta kuin yksinkertainen käytännössä.

Miksi tietoa anonymisoidaan?

Aluksi tulisikin kysyä, onko anonymisoinnissa ylipäätään mieltä. Puolestapuhujat voivat perustella anonymisointia mm. seuraavilla argumenteilla:

- Kun tiedot on anonymisoitu, ne voidaan jakaa vapaasti ja tutkijat voivat hyödyntää niiden sisältämiä tietoja.
- Vastaamme GDPR:n vaatimuksiin käsittelemällä vain anonymisoituja tietoja.
- Anonymisoituja tietoja voi käsitellä ilman suostumuksia.
- Voimme säilyttää, uudelleenkäyttää ja myydä anonymisoituja tietoja miten haluamme.

Kaikki argumentit ovat sellaisenaan täysin valideja, mutta asialla on myös kääntöpuolensa. Ei kannattajilla, vaan myös esimerkiksi tällaisia väitteitä kantansa tueksi:

- Vaikka anonymisoitu tieto olisikin anonyymia omassa kontekstissaan, se voidaan hyvin todennäköisesti de-anonymisoida yhdistelemällä tietoa muista tietolähteistä.
- Tutkijat eivät hyödy mitään siitä, jos he saavat käsiinsä täysin anonymisoitua tietoa.
- Miksi lähtökohtaisesti julkiseksi tarkoitettu (esim. Twitter tai keskustelupalsta) data pitäisi anonymisoida?
- Haluaako mikään taho ostaa tietoja, jos niistä on poistettu kaikki epäsuoratkin henkilötiedot, koska jäljelle ei jää mitään demografisia tietoja.

Olemme meneillään olevan Digitaaliset aineistot käyttöön -hankkeemme aikana kehitelleet täysin automaattista ratkaisua pdf-tiedostojen anonymisointiin.

Myös kaikki ei-argumentit ovat mielestäni sellaisenaan täysin valideja, joten anonymisoinnin tarvetta tulisikin tarkastella tapauskohtaisesti tarve- ja hyötylähtöisesti. Eli onko vaiva ja riski siis hyötyjen arvoista?

Toteutuessaan oikein, anonymisoituja tietoja ei pitäisi voida yhdistää henkilöön edes käyttämällä apuna muita tietolähteitä. Käytännössä tämä on erittäin vaikeaa, ellei jopa mahdotonta toteuttaa. Oletetaanpa, että olisin kirjoittanut nimettömänä jonkin tiettyä uskonnollista ryhmää kritisoivan kirjoituksen ja se olisi julkaistu lehdessä. Kirjoituksessa mainittaisiin kuitenkin tutkintonimikkeeni, asuinalueeni sekä ohimennen jotain lapsieni harrastuksista. Yhdistelemällä tietoja erinäisistä julkisista tietolähteistä, verkosta ja sosiaalisesta mediasta pystyttäisiin näiden kolmen tekijän yhdistelmällä rajaamaan vaihtoehdot todennäköisesti suoraan minuun tai korkeintaan muutaman ihmisen joukkoon.

Digitalian kehittämä ratkaisu anonymisointiin

Xamkin Digitalia-tutkimuskeskus on anonymisoinnin kannalla siitäkin huolimatta, että pieleen mennyt tai purettu anonymisointi voi johtaa yksityisyyden vaarantumiseen . Olemme meneillään olevan Digitaaliset aineistot käyttöön -hankkeemme aikana kehitelleet täysin automaattista ratkaisua pdf-tiedostojen anonymisointiin.

Jos lähtötilanne on esimerkiksi pdf-tiedosto, joka sisältää ainoastaan kuvina skannattua tekstiä, ratkaisumme ensimmäinen vaihe on OCR-luku (Optical Character Recognition) Tesseract-ohjelmalla. OCR tunnistettu teksti analysoidaan Polyglot NER (Named Entity Recognition) -työkalulla sekä hyödyntäen useita erilaisia sanalistoja.

Lopuksi ratkaisumme piirtää tunnistettujen anonymisoitavien kohteiden päälle sopivan kokoisen laatikon. Pdf-tiedostojen sisäiseen rakenteeseen perehtyneet tietävät, että pelkkä laatikko ei riitä alla olevan OCR-tiedon hävittämiseen. Tämän vuoksi viimeisin vaihe ratkaisussamme on "laatikoidun" pdf:n tallentaminen jälleen kuviksi ja näin luotujen kuvien OCR-luenta uudelleen. Koko ratkaisu on rakennettu toimivaksi Ubuntu Linux -järjestelmän päällä. Se perustuu avoimeen lähdekoodiin, vapaasti saatavilla oleviin ohjelmistoihin ja omaan Python-kehitykseemme.

Vaikka ratkaisussa on vaiheita suhteellisen monta, yksi sivullinen tekstiä pyörähtää OCR-luetuksi anonymisoiduksi sivuksi noin 30 sekunnin kuluessa ja ratkaisu on helposti skaalattavissa käyttämään kaikkia mahdollisia CPU-threadeja rinnakkain.

Jonkinasteinen anonymisointi, etenkin täysin automatisoitu sellainen, on hyvä ensimmäinen askel kohti vapaammin hyödynnettävää tietoa kontekstista riippumatta. Esimerkiksi arkistoissa olevat henkilötietoja sisältävät dokumentit tai median edustajien seulotut sähköpostit voitaisiin avata avoimeen käyttöön anonymisoinnin jälkeen. Ratkaisumme koodit ovat vapaasti saatavilla GitHubissa[4], joten kaikkien sisäiset nörtit liikkeelle ja kokeilemaan. Palautteita toimivuudesta otetaan Digitaliassa mieluusti vastaan.

Jonkinasteinen anonymisointi, etenkin täysin automatisoitu sellainen, on hyvä ensimmäinen askel kohti vapaammin hyödynnettävää tietoa kontekstista riippumatta.

TkT Anssi Jääskeläinen toimii TKI-asiantuntijana Kaakkois-Suomen ammattikorkeakoulussa ja vastaa Digitalia-tutkimuskeskuksessa teknisestä kehityksestä. www.digitalia.fi

LÄHTEET

1. *https://tietosuoja.fi/pseudonymisointi-ja-anonymisointi*
2. *https://www.tietosuojatyokalu.fi/ohjeet/sanasto/*
3. *Narayanan, A & Shmatikov, V. (2008). 'Robust De-anonymization of Large Sparce Datasets'. SP 2008, 111–125.*
4. *https://github.com/Digitalia-Xamk/python-anonymizer*