Omadata osana AuroraAI-hanketta

Laatinut: Mika Honkanen (mika.honkanen@vrk.fi)

Työpaketti neljässä keskityttiin rakentamaan pienin toimiva tuote (engl. MVP, Minimum Viable Product). Osana työpakettia pohdittiin alustavasti myös karkealla tasolla omadatan (engl. MyData) tuomista osaksi Auroran ohjelmistoarkkitehtuuria. Tarkempi omadatan viitearkkitehtuuri julkiseen hallintoon on laadittu erikseen osana yhteinen tiedonhallinta (YTI) -hankkeen työpaketti ykköstä[1]. Liitettä lukiessa kannattaa huomioida, että sekä omadata että tekoäly ovat tällä hetkellä voimakkaasti kehittyviä ja nopeasti muuttuvia asioita.

Aurora on “älykkäiden” palveluiden avoin ja hajautettu verkosto. Tavoitetilassa ihminen voi luvittaa omia henkilötietojaan itse Aurora -tekoälyverkon hyödynnettäväksi erilaisissa käyttötapauksissa. Intensio-analyysi hyödyntää käyttäjän luvittamia luotettavia attribuutteja esimerkiksi julkisen hallinnon tietovarannoista. Attribuutteja ovat esimerkiksi etunimi, sukunimi, osoite jne. Intensio-analyysin luotettavuutta ja attribuuttien rikkautta voidaan parantaa omadatan avulla rikastamalla ja automatisoimalla käyttäjän syötteitä. Näin käyttäjän ei tarvitse itse syöttää samoja tietoja uudestaan esimerkiksi näppäimistön tai puhe-käyttöliittymän avulla ja toisaalta tiedot tulevat syötettyä aina oikein (käyttäjän syötteessä ei ole virheitä tai niiden merkitystä ei tarvitse tulkita). Valtion digitalisoinnin periaatteissa on linjattu, että “pyydämme uutta tietoa vain kerran”, “hyödynnämme jo olemassa olevia julkisia ja yksityisiä sähköisiä palveluita” ja “avaamme tiedon ja rajapinnat yrityksille ja kansalaisille” [2]. Nämä tukevat jo olemassa olevan omadatan käyttöä osana Auroraa.

Omadatan (engl. MyData) ideana on antaa ihmisen itsensä päättää omien henkilötietojensa käytöstä[3]. Käytännössä ihmiset myöntävät ”suostumuksia” eli käyttölupia eri toimijoiden välisiin henkilötietoja käsitteleviin tietovirtoihin. Tekoälyn avulla voidaan tarjota palveluita usein aiempaa helpommin. Jotta ihmisiä käsittelevä tiedonhallinta olisi ihmislähtöistä, pitää seuraavien asioiden toteutua:

1. Kaikki verkostoon liitettävät tietovarannot ovat riittävän hyvin kuvattu (semanttisesti riittävän yhteentoimivia)
2. Tietovarantoja pystyy hyödyntää kehittäjäystävällisten rajapintojen avulla. Tietovarantoihin on rakennettu tarvittavat rajapinnat. Rajapintoihin voidaan toki rakentaa myös erilaisia sovittimia, mutta yleensä ne tekevät ohjelmistoarkkitehtuurista tarpeettoman monimutkaista ja lisäävät turhia kustannuksia. Yksinkertaisuus on tässäkin sekä kustannustehokasta että ohjelmistoarkkitehtuurin kannalta kaunista
3. Ihmisten antamat suostumukset on koottu suostumusrekisteriin koneluettavassa muodossa
4. Aurora-verkosto pyytää suostumuksia kansalaiselta niin, että kansalainen tietää yksiselitteisesti mitä tietoja ja kenelle niitä ollaan luvittamassa. Verkosto koostaa suostumusta varten attribuutti-listan ja tietoja käsittelevät osapuolet. Attribuutit ovat semanttisesti yhteentoimivia tietovarantojen kuvaamisessa hyödynnettävän metadatan kanssa

[1] http://goo.gl/NN7d41
[2] https://vm.fi/digitalisoinnin-periaatteet
[3] http://julkaisut.valtioneuvosto.fi/handle/10024/160954

Arkkitehtuurin kuvailemassa palvelurekisterissä (palvelun nimi | asiasanat | palvelun osoite) on paljon samaa Suomi.fi-palvelutietovarannossa (lyhyesti PTV:ssä). PTV:ssä on kuvattu palvelut, asiasanat ja niihin liittyvät asiointikanavat. Kannattaisi selvittää, voisiko PTV:tä jatkokehittää lisäämällä siihen uuden asiointikanavan, rajapinta (API), kuvatulle palvelulle nykyisten puhelimen, sähköpostin, käyntiosoitteet ja internet-sivun rinnalle. PTV tietomalli on kuvattu tarkemmin täällä: https://tietomallit.suomi.fi/model/suomiptv/.

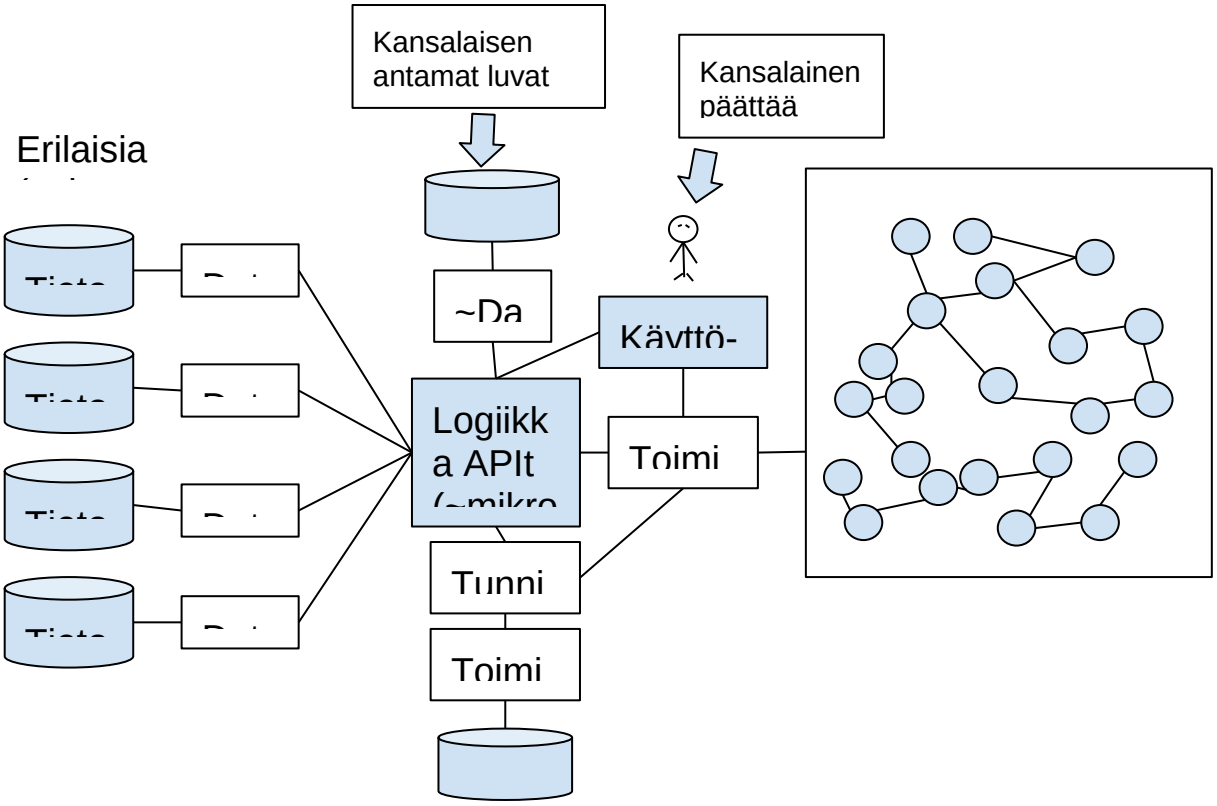
Datan anonymiteetti ja tietosuoja ovat selkeä positiivinen haaste koko Aurora verkoston toiminnan kannalta. Attribuuteista muodostuu helposti melko luotettava sähköinen identiteetti. Jos esimerkiksi tiedetään, että joku on ollut kello 16:30 Helsingissä 7 ratikassa (yksi attribuutti), pystytään rajaamaan päättely siitä, ketä henkilöä kyseinen data koskee noin 5,5 miljoonasta henkilöstä (Suomen väkiluku) alle 60 henkilöön (kaikki raitiovaunussa silloin olleet henkilöt). Muutamalla attribuutilla saatetaan pystyä päättelemään henkilön identiteetti hyvin luotettavasti riippuen attribuuteista ja niiden luotettavuudesta. Esimerkiksi jos sama henkilö on käynyt kaupassa (esimerkiksi S-marketissa) tämän jälkeen jne.

HUS on esimerkiksi arvioinut, ettei se käytännössä pysty tekemään datan anonymisointia terveystiedoista, koska Suomen väestö on siihen liian pieni populaatio ja tarkkojen terveydentilaan kuvaavien attribuuttien avulla on liian helppo päätellä ja purkaa datan anonymiteetti. Esimerkiksi avoimen datan anonymisointi on pystytty purkamaan aika innovatiivisin keinoin maailmalla. New Yorkissa yhdistettiin anonymisoitua avointa dataa takseista ja videokameroiden kuvia, jolloin datan anonymiteetti onnistuttiin purkamaan[4]. Mikäli Aurora sisältää historiatietoa identiteeteistä ja attribuuteista, on hankalaa säilyttää datassa anonymiteetti. Asiaa voi yrittää esimerkiksi summaamalla dataa kategorioihin (esimerkiksi Stiglitzin malli[5]). Tärkeintä on tiedostaa ja huomioida osana hyvää suunnittelua tämä asia.

Vaihtoehtoinen toteutustapa on se, että Aurora verkosto tuottaa osaksi omadata-suostumusta tiedon siitä, mitä dataa ja mille palveluille se aikoo välittää (esimerkiksi attribuutti-listaus ja sen vastaanottavat osapuolet & palvelut). Tässä tapauksessa verkoston tulisi aina paljastaa sen dynaaminen rakenne siinä vaiheessa, kun kansalainen kertoo itsestään sille ihmiskeskeisesti henkilötietojaan (attribuutteja).

Kuvassa 1 on tuotu alustavasti Aurora ja omadatan ideat yhteen.

[4] https://dash.harvard.edu/bitstream/handle/1/30340010/OpenDataPrivacy.pdf
[5] http://www.stakes.fi/yp/2009/5/simpura.pdf



Kuva 1 on yksi tapa esittää asia. Se yksinkertaistaa vähän asioita. Esimerkiksi melkein kaikki tekoäly-toteutukset toimivat rajapinnan tai rajapintojen avulla ja niitä voi liittää useisiin kohtiin tekemään erilaisia asioita. Tämän lisäksi Aurora-verkossa voidaan siirtää myös dataa eri tavoin. Esimerkiksi siirtämällä osoitinta dataan (esimerkiksi hyperlinkki) ja pääsy-avainta (salauksen purkamis-avain dataan). On tärkeää ymmärtää, että rajapintoja on paljon erilaisia. Karkeasti voidaan ajatella, että toiset tarjoavat dataa ja toiset tarjoavat sen lisäksi sen käsittelyä. Rajapinta-lähtöinen tapa suunnitella ohjelmistoja tarkoittaa, että mikä tahansa suuri tai pieni kokonaisuus pilkotaan pieniin paloihin, joista jokainen toteutetaan rajapintojen avulla. Suurinkin monoliitti-ohjelmisto (elefantti) syödään siis pieni pala (rajapinta) kerrallaan. Rajapintoja voi lisäksi käyttää yhä uudestaan. Jolloin mitään asiaa ei periaatteessa tarvitse tehdä kuin kerran. Käytännössä asioita iteroidaan useita kertoja ja sama rajapinta saatetaan toteuttaa useilla eri ohjelmointikielillä, jolla saadaan suorituskykyeroja (toisella ohjelmointikielellä toteutettu voi olla nopeampi, luotettavampi jne.)

Ihmisten suostumukset eli ihmisten antamat käyttöluvat omien tietojen käsittelyyn on tallennettu koneluettavassa muodossa suostumusrekisteriin, joka voi olla hajautettu (useita eri tietovarantoja) tai keskitetty (yksi tietovaranto). Pääasia on, että sen rajapintojen rakenne (esimerkiksi Open API) ja tietomallit (suostumukset ja suostumusrekisteri) ovat standardeja ja ne on yhdessä sovittu. Lisäksi täytyy sopia tietosuojaan liittyvistä asioista. Näitä ovat esimerkiksi erilaiset pääsynhallinta-asiat. Verkostossa voi siis olla useita suostumusrekistereitä, eikä ne vaikuta käyttäjäkokemukseen mitenkään, jos asia vaan suunnitellaan oikein. Rajapintoja Auroran kaltaisessa verkossa on helposti tuhansia, joten niiden toteuttamiseen tarvitaan samankaltaiset suunnitteluperiaatteet, johon kaikki osapuolet sitoutuvat etukäteen. Ja mahdollisesti muuta arkkitehtuuria.

Tämän lisäksi arkkitehtuurisuunnittelussa on hyvä huomioida EU:n tietosuoja-asetus[6], julkisuuslaki[7], hallintolaki[8] ja se, että Suomessa henkilötietojen käytöstä säädetään arviolta noin 700-800 eri laissa. Lainsäädännön suunnittelu on tähän asti perustunut pitkälti (1800-luvun alun Ranskan vallankumous) siihen ajatukseen, että lakien avulla valtion ja julkisen hallinnon toimintaa ja vallankäyttöä pyritään rajoittamaan kansalaisiin nähden. Siksi lainsäädännön avulla on pyritty asettamaan rajoja siihen, mitä tietoja viranomaiset saavat kansalaisesta kerätä ja mihin käyttötarkoitukseen kerättyä tietoa saa hyödyntää. Tässä ajattelussa ei olla kuitenkaan huomioitu teknologian kehittymistä ja sitä ajatusta, että myös kansalainen voi olla aktiivinen toimija omien henkilötietojensa käsittelyssä. Eli sitä ajatusta, että kansalaisella olisi oikeus päättää (ainakin jossain määrin) omien tietojensa käytöstä esimerkiksi Aurorassa.

Loppuun on hyvä todeta, että hyvä suunnittelu on myös sellaista, ettei kenenkään tarvitse -- jos ei itse halua -- alkaa luvittamaan omia tietojaan. Luodaan siis mahdollisuuksia, mutta ei esteitä tai rajoitteita uuden teknologian avulla.

[6] https://eur-lex.europa.eu/legal-content/FI/TXT/?uri=CELEX%3A32016R0679
[7] https://www.finlex.fi/fi/laki/ajantasa/1999/19990621
[8] https://www.finlex.fi/fi/laki/ajantasa/2003/20030434