

Invinite Insights:

**SEMANTTINEN TIETOPOHJA
RATKAISEE TERVEYDENHUOLLON
TEKOÄLYN SKAALAUTUMISEN**

**DATA AS SOFTWARE –
OHJELMISTOKEHITYSLÄHTÖINEN TIETOTUOTANTO
TERVEYDENHUOLLOSSA**

**Ville Siltala
Severi Niskanen
Antti Brunni**

Tiivistelmä

Terveystieteenhuollossa kehitetään parhaillaan tekoälyyn pohjautuvia ratkaisuja monenlaisiin tarpeisiin: diagnostiikka-algoritmeja, ennustavaa analytiikkaa ja kliinisiä päätöksentekijärjestelmiä. Monen hankkeen upea tavoite jää kuitenkin toteutumatta kehittämisen jäädessä pilottivaiheeseen. Usein syy ei ole teknologiassa tai tekoälymallissa vaan tietopohjassa: data ei ole riittävän yhdenmukaista, jäljitettävää, kliinisesti luotettavaa tai sitä ei ole saatavilla.

Nykyisessä tietotuotannossa semanttinen harmonisointi, testaus, dokumentaatio, versionhallinta ja omistajuus jäävät usein liian heikoiksi. Tieto voi näyttää teknisesti käyttökelpoiselta, vaikka sen laatu ja rakenteellinen yhdenmukaisuus ovat epävarmoja. Tämä tekee esimerkiksi tekoälyratkaisusta hauraita: malli voi toimia pilotissa, mutta epäonnistua tuotannossa, jos koulutusaineisto ei perustu yhtenäiseen ja hallittuun tietopohjaan.

Ratkaisuksi tarvitaan ajattelua, jossa datantuotantoa johdetaan samalla kurinalaisuudella kuin kriittistä ohjelmistokehitystä. Tämä tarkoittaa versionhallintaa, automaattista testausta, datakontrakteja, CI/CD-käytäntöjä, selkeitä vastuuta ja data-arkkitehtuuriin rakennettua semanttista harmonisointikerrosta. Niin kutsuttu datan mallinnus -kerros on tarve nostaa keskeiseksi: siinä heterogeeninen lähdedata muutetaan yhtenäiseksi, kliinisesti mielekkäiksi käsitteiksi ennen analytiikka- ja tekoälykäyttöä.

Suomella on vahva lähtökohta terveystieteen hyödyntämiseen, mutta poikkeuksellisen laadukkaat datavarannot eivät yksin riitä. Tavoite tietopohjan semanttisen laadun kehittämisestä on rakennettava pysyväksi osaksi arkkitehtuureja, hankintoja ja tietotuotannon johtamista. Vasta sen jälkeen tekoälystä voidaan odottaa luotettavia, tuotantokelpoisia ja laajasti skaalautuvia hyötyjä terveydenhuollossa.

Taustatiedot ja havainnot

Terveysdata kompleksisessa ja muuttuvassa toimintaympäristössä

Terveydenhuollon tietojärjestelmä ja tietojohdamiskenttä on poikkeuksellisen heterogeeninen. Yksittäisellä hyvinvointialueella voi olla samanaikaisesti käytössä useita potilastietojärjestelmiä ja kymmeniä muita toimintaa ohjaavia tietojärjestelmiä, joilla on omat käsitteensä, tietomallinsa, koodistonsa, kirjauskäytäntönsä ja rajapintansa. Niiden yhdistäminen analyyttisesti käyttökelpoiseksi tietoaaineistoksi moninaisiin ja monitahoisiin käyttötarpeisiin, huomioiden sekä paikallinen toiminta että kansalliset koodistot, ei ole kertaluonteinen hanke vaan jatkuvaa tuotantotyötä.

Kun valmiiksi kompleksiin ympäristöön yhdistyvät hyvinvointialueilla käynnissä olevat potilastietojärjestelmien konsolidoinnit ja järjestelmien suuret versiopäivitykset, palvelutuotannon uudelleenorganisointumiset ja toimintatapamuutokset sekä masterdata-järjestelmien puuttuminen, aiheuttaa tämä alueiden tietotuotannoille erittäin korkeat vaatimukset muutoshallinnalle ja jatkuvalle työn validoinnille. Tähän haasteeseen nykyiset tekniset toteutukset ja toimintamallit eivät ole pystyneet vastaamaan.

Terveysdatan rakenne on semanttisesti kompleksinen

Terveysdatan semanttinen monimutkaisuus on laadullisesti erilainen haaste kuin esimerkiksi talous- tai logistiikkadatassa. Sama kliininen käsite voi esiintyä ICD-10-diagnoosikoodina, SNOMED CT -viitekehyksen mukaisena käsitteenä, vapaatekstinä, lääkitystietojärjestelmän omana luokitteluna tai laboratorioarvona LOINC-koodistossa. Siksi järjestelmien tekninen integraatio ei yksin riitä. Se, että tieto siirtyy järjestelmästä toiseen, ei vielä takaa, että se säilyttää merkityksensä.

Tarvitaan semanttista integraatiota: yhteinen ymmärrys siitä, mitä tieto tarkoittaa eri järjestelmissä ja käyttötarkoituksissa. Tutkimuskirjallisuus tukee tätä havaintoa. Terveysdatan laadun ylläpito on pysyvä operatiivinen haaste (Goldstein ym. 2024), ja samalla skaalautuvat dataputket on tunnistettu keskeiseksi edellytykseksi tekoälyratkaisujen hyödyntämiselle terveydenhuollossa (Syed ym. 2024). Terveydenhuollon tietovarastojen rakentaminen edellyttää standardoituja prosesseja, ei pelkkiä teknologiahankintoja (Häyrynen ym. 2024).

Nykytila: dataa tuotetaan käsityönä

Suomalaisissa terveydenhuollon organisaatioissa data liikkuu usein järjestelmien välillä prosesseissa, joita ei ole riittävästi dokumentoitu, versioitu, testattu tai valvottu. Muunnoksia ajetaan käsin tai puoliautomaattisilla komentosarjoilla, joiden ylläpito voi olla yksittäisten asiantuntijoiden varassa.

Kun tekijä vaihtaa tehtävää tai lähtee organisaatiosta, osa tuotantologiikasta voi kadota hänen mukanaan. Ongelmaa vahvistaa se, että käsittelyn toimintalogiikka on usein hajautettu eri raportointi- tai käyttötapauskohtaisiin dataputkiin sen sijaan, että se olisi keskitetty yhteen hallittuun ratkaisuun. Tällöin muutokset joudutaan tekemään erikseen useisiin kohteisiin, mikä sitoo kehittämiseen varattuja resursseja nykyisen tietotuotannon jatkuvaan ylläpitoon.

Tämän seurauksena datassa piilevä totuus on ripoteltu tiedonmurusina useisiin erilaisiin raportointisovelluksiin eikä niistä ole saatavilla yhtä keskitettyä tietovarastoa, josta jokainen käyttötapaus johdetaan. Yhdet luvut, yksi totuus –periaate ei toteudu.

Dataputkia tai niiden komponentteja ei testata samalla tavalla kuin ohjelmistoja. Yksikkö- ja integraatiotestit puuttuvat, ja laaduntarkistukset tehdään usein vasta lopputuloksesta visuaalisesti käyttäjän toimesta nojaten täysin manuaaliseen resursseja hukkaavaan prosessiin. Semanttinen harmonisointilogiikka voi olla hajallaan kehittäjien muistissa, dokumenttien reunoilla tai erillisissä Excel-taulukoissa. Tilanne muistuttaa ohjelmistokehitystä ilman versionhallintaa, testausta ja dokumentaatiota. Ohjelmistokehityksessä tällainen toimintamalli tunnistettiin kestäättömäksi jo kauan sitten. Datantuotannossa se on silti monin paikoin edelleen arkea.

Markkinatilanne: yleistyökalut eivät ratkaise terveystietojen semanttisia ongelmia

Markkinoilla on useita yleisiä tietovarastotyökaluja ja -alustoja, jotka perustuvat SQL-transformaatioihin ja pilvi-infrastruktuuriin. Ne tarjoavat osittaisen ratkaisun: ne standardoivat transformaatiokerrosta, mutta eivät sisällä terveystietojen ominaista semanttista mallinnusmenetelmää, tarpeen vaativaa muutosjoustavuutta tai pohjaa datan ja käsittelykoodin laadun varmistavalle automaattiselle testaukselle.

Pelkästään SQL-pohjaisilla lähestymistavoilla ei voida luotettavasti hallita terveysdatan semanttista monimutkaisuutta. Haasteita syntyy esimerkiksi koodistoriippuvuuksista, aikaviittauksista, potilasidentiteetin yhdistämisestä usean lähteen yli sekä potilasturvallisuuden kannalta kriittisistä validointisäännöistä. Jos nämä ongelmat jäävät ratkaisematta datantuotannon vaiheessa, ne eivät katoa tekoälyratkaisuihin siirryttäessä vaan alkavat tuottaa harhaanjohtavia tuloksia.

Uudet paradigmat: data tuotteena ja tekoälyn syötteenä

Kansainvälisessä tutkimuksessa datan omistajuuden selkeyttäminen ja data-as-a-product-ajattelu ovat nousseet keskeisiksi organisatorisiksi periaatteiksi (Azeroual & Theel 2023; Schubotz ym. 2024). Data mesh -arkkitehtuuri, hajautettu omistajuus ja datakontraktit siirtävät vastuun datan laadusta lähemmäs niitä toimintoja, joissa data syntyy.

Tekoälyn näkökulmasta tämä on ratkaiseva muutos. Korkealaatuinen ja semanttisesti yhtenäinen datatuote on tekoälymallin tärkein yksittäinen edellytys. Se on usein tärkeämpi kuin mallin arkkitehtuuri tai käytettävissä oleva laskentakapasiteetti. Jatkuvasti muuttuvassa ympäristössä ainut mahdollisuus on tukeutua käsitepohjaiseen malliin, jossa yhteisesti käytetty tietopohja voidaan korjata ympäristöstä aiheutuneiden muutosten johdosta vain kerran, ei hajautetusti raporttikohtaisine dataputkineen.

Ongelman kuvaus

Terveysdatan ongelmien juurisyy ei useimmiten ole lähdejärjestelmien datassa, vaan siinä, miten datantuotantoa johdetaan. Kun dataputkia ei käsitellä insinöörityönä ja semanttinen harmonisointilogiikka jää arkkitehtuurissa näkymättömäksi, vastuut, testaus, versionhallinta ja dokumentaatio hajoavat organisaation eri osiin tai jäävät toteutumatta. Tämän seurauksena tieto voi näyttää teknisesti käyttökelpoiselta, mutta sen kliininen merkitys, jäljitettävyyys ja luotettavuus jäävät epävarmoiksi.

Juurisyy: datatuotantoa ei johdeta insinöörityönä eikä semantiikkaa hallita arkkitehtuurissa

Terveysdatan ongelmat eivät useimmiten johdu siitä, että data olisi lähtökohtaisesti puutteellista lähdejärjestelmissä. Ne johtuvat siitä, että datan tuotantoprosesseja ei johdeta insinöörityön standardien mukaisesti. Erityisenä ongelmana on, että semanttinen harmonisointilogiikka ei ole arkkitehtuurissa näkyvä, hallittu ja omistettu kokonaisuus. Keskeisiä juurisyitä ovat seuraavat.

Omistajuuden puuttuminen.

Kukaan ei vastaa dataputken kokonaisuudesta. Tekniset osiot kuuluvat IT:lle, sisältö substanssiasiantuntijoille ja laatu ”kaikille” — eli käytännössä ei kenellekään.

Tyypillisesti omistajuus ja vastuu puuttuu erityisesti semanttisen harmonisointilogiikan osalta. Tyypillisesti myös suuri osa integraatioista ja käsittelytyöstä on ulkoistettu ulkoiselle toimijalle, joka ei toimi saumattomasti yhteen muun tietotuotannon kanssa.

Testauksen puuttuminen.

Dataputkissa ei käytetä yksikkötestejä, integraatiotestejä, datakontraktien validointia eikä hyväksyntätestejä. Semanttiset virheet havaitaan usein vasta analyysikerroksen tuloksissa: kaksi eri käsitettä on niputettu yhdeksi tai yksi käsite on hajonnut useaksi eri tulkinnaksi. Pahimmillaan virhe huomataan vasta, kun malli tuottaa kliinisesti epäuskottavan suosituksen. Testauksen puuttuessa muutoksia ei uskalleta tehdä monikäyttöisiin kokonaisuuksiin vaan ajaudutaan tekemään hajautettuja dataputkia erillISRatkaisuina.

Versionhallinnan puuttuminen.

Muutoksia tehdään suoraan tuotantoympäristöön ilman hallittua muutosprosessia. Semanttinen muutos, esimerkiksi diagnoosiluokittelun päivitys, voi levitä dataputkeen hiljaisesti ilman muutoshistoriaa tai tiedotusta. Tällöin muutosten jäljitettävyyks ja auditointi vaikeutuvat olennaisesti.

Dokumentaation puuttuminen.

Transformaatiologiikka voi olla yksittäisen kehittäjän muistissa tai hajallaan erillisissä tiedostoissa. Semanttinen tieto — esimerkiksi päätös siitä, että DM2 ja E11 yhdistetään tietyssä käyttötapauksessa samaksi käsitteeksi — katoaa helposti organisaatiosta hiljaisen tiedon mukana.

Ympäristöjen ajautuminen erilleen.

Kehitys-, testi- ja tuotantoympäristöt eroavat toisistaan tavoilla, joita ei hallita systemaattisesti. Tällöin semanttisesti koulutettua mallia voidaan käyttää tuotantoympäristössä, jossa lähdedata ei enää vastaa koulutushetken aineistoa. Tämä heikentää tietomallin luotettavuutta ja vaikeuttaa tulosten arviointia. Varsinkin datapohjaisessa tekoälykehityksessä kehitysympäristön tulisi olla tiiviisti yhteydessä ja samalla alustalla varsinaiseen tietotuotannon kanssa, jotta kehitetyt ratkaisut voidaan myös implementoida ketterässä aikataulussa.

Operointiprosessien ja -työkalujen puute.

Datatuotteiden kehitys nojaa usein manuaalisiin prosesseihin, jossa kehitysversiot siirretään manuaalisesti ympäristöstä toiseen. Automaation ja prosessien kehityksen puute johtaa tilaan, jossa datatuotteiden elinkaari on epäselvä. Ilman kehittynyttä DevOps-prosessia on vaikea tunnistaa, mikä koodiversio tulisi olla ajossa missäkin ympäristössä ja siten, mikä data on tuotettu milläkin koodilla. Vikatilanteiden havainnointi jätetään myös manuaalisen työn varaan. Datassa ajan yli tapahtuvien muutosten aiheuttamat virheet päätyvät loppukäyttäjien havaittavaksi tai pahimmassa tapauksessa niitä ei havaita ollenkaan.

Semanttisen kerroksen puuttuminen.

Jos yksi potilastietojärjestelmä kirjaa tyypin 2 diabeteksen koodilla E11, toinen koodilla DM2 ja kolmas vapaatekstimerkinnällä t2dm, kyse voi kliinisesti olla samasta asiasta. Ilman semanttista harmonisointia tiedon hyödyntäjä tai tekoälymalli ei kuitenkaan käsittele näitä yhtenä käsitteenä, vaan erillisinä muuttujina. Tällöin tietomalli voi aliarvioida diagnoosin esiintyvyyden ja voidaan päätyä tulkintaan jossa kirjaamiskäytäntöjen muutokset näyttäytyvät kliinisinä muutoksina.

Ratkaisut ja suositukset

Ratkaisuna kuvattuihin ongelmiin pystytään vastaamaan

- 1.ohjelmistolähtöisellä datatyöllä
- 2.käsitepohjaisella yhteisellä tietopohjalla ja semanttisella kerroksella
- 3.datatuote ja tarvelähtöisellä datatyön johtamisella

Data as Software -menetelmä tuo datan tuotantoprosessit, tietoputket, muunnokset ja semanttisen harmonisoinnin samanlaisen hallinnan piiriin kuin kriittiset ohjelmistojärjestelmät. Käytännössä tämä tarkoittaa versionhallintaa, automaatiotestausta, CI/CD-putkistoja, datakontrakteja ja selkeästi määriteltyä omistajuutta.

Tietotuotannossa on pyrittävä väsymättömästi ja pitkäjänteisesti muuttamaan heterogeeninen lähdedata yhtenäisiksi ja kliinisesti mielekkäiksi käsitteiksi ja rakentamaan näihin pohjautuvaa semanttista kerrosta.

Tietotuotantoa on johdettava järjestelmällisesti kokonaisuutena koko tietotuotannon jalostusputken osalta sekä vahvasti tarvelähtöisesti.

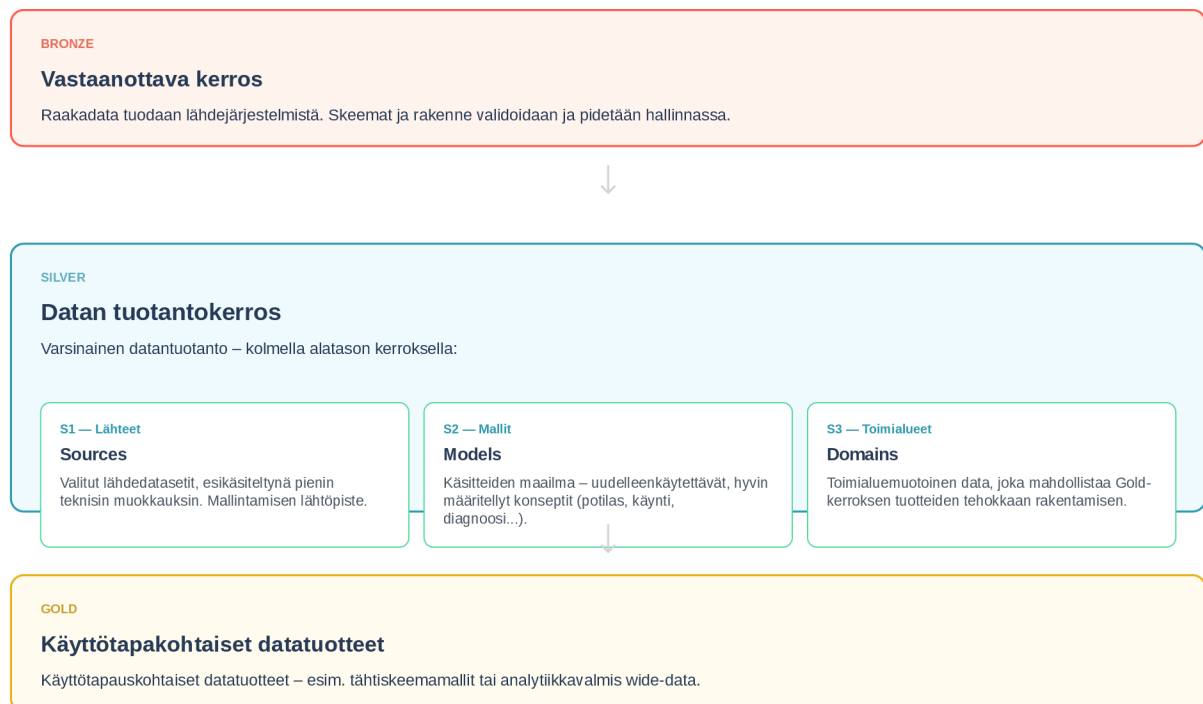
Tietotuotannon johtamisessa korostuu näkyvyys ja ymmärrys koko tietotuotantoputken työtehtäviin ja tuotoksiin, ei pelkästään käyttökerroksen tarpeisiin. Vastuut ja omistajuudet on hyvä jakaa käsitteellisten tietokokonaisuuksien mukaan. Se tarkoittaa, että täytyy aktiivisesti irrottautumista lähdejärjestelmien tai raportointitarpeiden määrittämästä perustasta.

Systemaattinen viitekehys: Mitaliarkkitehtuuri ja semanttinen harmonisointikerros

Tietotuotannon ongelmia ei ratkaista yksittäisillä työkaluhankinnoilla ja toistamalla vanhoja toimintamalleja uudessa työkalussa. Erityisesti uutena käyttöön otettavat tietoallaspohjaiset ratkaisut ovat alttiita vanhojen syntien toistamiselle, tarjotessaan selkeästi aiempia järjestelmiä löyhemmän kehikon käytännön datatyön tekemiselle.

Tämä mahdollistaa vanhan järjestelmän asiantuntijoille luontevan väylän jatkaa kehittämistä totuttujen tapojen mukaisesti. Yhtälöön tarvitaan siis tiukemmin määritelty uusi arkkitehtuurinen viitekehys, joka kattaa tuotantoketjun raakadatasta käyttötapaukseen. Tähän vastaamaan on kehitetty Data as Software -menetelmä.

Mitaliarkkitehtuuri jäsentää tämän pronssi-, hopea- ja kultakerrokseen seuraavasti:



Kaavio: Data-as-Software tietotuotannon rakenne

Semanttinen taso tekoälyn kriittisenä edellytyksenä

Tekoälyn tuotantokäytön suurin este terveydenhuollossa ei useinkaan ole mallien tekninen laatu, vaan koulutusaineiston semanttinen epäyhtenäisyys. Kyse ei ole vain datan puuttumisesta tai teknisestä laadusta, vaan siitä, tarkoittavatko eri lähteistä tulevat tiedot samaa asiaa.

Tätä ongelmaa ei ratkaista vain lisäämällä laskentatehoa tai vaihtamalla tietotuotanto allaspohjaiseen malliarkkitehtuuriin. Se ratkaistaan rakentamalla arkkitehtuuriin semanttinen harmonisointikerros ennen kuin dataa käytetään tekoälymallin kouluttamiseen. Mitaliarkkitehtuurissa hopeakerros on tässä keskeinen: sen tehtävänä on muuntaa eri lähteiden käsitteet yhtenäisiksi ja kliinisesti mielekkäiksi tietokokonaisuuksiksi.

Tekoälyn tuotantokäyttö edellyttää lisäksi enemmän kuin onnistunutta pilottia. Tarvitaan hallitut syötteet, määritelty julkaisulogiikka, mallin ajautumisen seuranta, selkeä omistajuus, integraatio operatiiviseen ympäristöön ja ylläpidettävä elinkaarimalli. Ilman näitä edellytyksiä pilotti ei muutu tuotantokelpoiseksi palveluksi. Data as Software -menetelmä rakentaa nämä edellytykset tietopohjan tasolle.

Suosituks

Suositus 1: Määritellään kansallinen tietotuotannon kypsyysmalli terveydenhuollolle

Tarvitaan kansallisesti määritelty kypsyysmallin terveystietojen tuotantoprosesseille. Mallin tulee asettaa vähimmäisvaatimukset versionhallinnalle, testaustilanteille, omistajuudelle, dokumentaatiolle, auditointitietojen ja semanttiselle harmonisoinnille.

Ilman yhteistä viitekehystä hyvinvointialueet rakentavat keskenään yhteensopimattomia ratkaisuja. Tämä ylläpitää semanttista epäyhtenäisyyttä ja kehitystyön jakamisen edellytyksiä alueiden välillä, kansallisten ohjaamista sekä kansallisten tekoälyhankkeiden pohjaa.

Kypsyysmalli loisi vertailukelpoisuutta, ohjaisi hankintoja ja parantaisi kansallisten rekisterien ja tilastojen käsitteiden yhtenäisyyttä. Ensimmäisenä askeleena käynnistetään kuuden kuukauden selvityshanke, jossa kartoitetaan nykyiset semanttisen harmonisoinnin käytännöt kahdella- kolmella hyvinvointialueella ja laaditaan kypsyysmallin ensimmäinen versio.

Suositus 2: Otetaan käyttöön semanttiset datakontraktit potilasturvallisuuskriittisissä tietovirroissa

Hyvinvointialueet ja järjestelmätoimittajat veloitetaan dokumentoimaan ja testaamaan datakontraktit potilasturvallisuuden kannalta kriittisissä tietovirroissa. Datakontrakti määrittelee skeeman ja sen metatiedot (semanttiset selitteet datalle), sallitut arvot, koodistot ja niiden versiot, laatuvaatimukset sekä muutoksenhallinnan menettelyt. Lääkitystiedot, diagnoosikoodit ja toimenpidekirjaukset kulkevat usein järjestelmien välillä ilman eksplisiittistä sopimusta tiedon merkityksestä.

Kun tällainen data päättyy tekoälymallin koulutusaineistoksi, malli voi oppia epäluotettavia käsitteitä. Ensimmäisenä askeleena pilotoidaan semanttinen datakontraktikäytäntö yhden hyvinvointialueen ja yhden potilastietojärjestelmätoimittajan välisessä integraatiossa vuoden 2026 aikana.

Suositus 3: Standardoidaan käsitteiden harmonisointikerros osaksi terveydenhuollon tietovarastoarkkitehtuuria

Terveydenhuollon tietoarkkitehtuuriin otetaan käyttöön kansallisesti yhtenäinen pronssi–hopea–kulta-kerrosmalli. Mallissa Silver kerroksen S2-taso vastaa semanttisesta harmonisoinnista: heterogeeniset lähdekohtaiset käsitteet muunnetaan yhtenäisiksi kliinisiksi käsiteyksiköiksi ennen analytiikka- ja tekoälykäyttöä.

Ilman käsitteiden systematisointia tekoälymalleja koulutetaan käsitteellisesti epäyhtenäisellä aineistolla. Semanttinen harmonisointikerros on kohta, jossa terveydenhuollon tekoälyvalmiuden keskeinen este — käsitekohtainen epäyhtenäisyys — ratkaistaan rakenteellisesti.

Ensimmäisenä askeleena julkaistaan suositusarkkitehtuurimalli ja avoin referenssi-implementaatio rajatulla tietosisällöllä (esimerkit diagnoosin, lääkityksen ja käynnin tietosisällöt)

Suositus 4: Sisällytetään semanttinen laatu ja datatuotannon insinööriosaaaminen hankintakriteereihin

Julkisten hankintojen kriteeristöä täydennetään vaatimuksilla, jotka koskevat datatuotannon insinööriosaaamista. Toimittajilta edellytetään näyttöä CI/CD-käytännöistä, automaattisesta testauksesta, versionhallinnasta, semanttisen harmonisointilogiikan hallinnasta ja dokumentoinnista.

Erityisesti tekoälyhankinnoissa tulee osoittaa, miten koulutusaineiston semanttinen laatu varmistetaan. Nykyiset hankintakriteerit painottuvat liikaa toiminnallisuuksiin ja lisenssimalleihin, jolloin tietopohjan laatu paljastuu ongelmaksi vasta käyttöönotossa.

Ensimmäisenä askeleena laaditaan tekoälyhankintoihin semanttisen datalaadun arviointikehikko, jota pilotoidaan seuraavassa laajamittaisessa terveydenhuollon tekoälyhankinnassa.

Suositus 5: Pilotoidaan Data as Software -menetelmään perustuva yhteiskehityshanke useamman hyvinvointialueen kesken ennen vuotta 2028

Erillisrahoituksella käynnistetään kaksi Data as Software -pilottia eri kokoisilla hyvinvointialueilla. Piloteissa toteutetaan täysi DaS-implementaatio: versionhallinta, testauskerrokset, semanttiset datakontraaktit, CI/CD-putkisto, pronssi–hopea–kulta-kerrosrakenne ja selkeä omistajuusmalli. Pilotissa yhteiskehitteisesti valmistuvat tuotokset jäävät osallistuvien alueiden käyttöön ja omistukseen, eivätkä muodostu osaksi yksittäisen teknisen toimittajan tuotetta.

Pilottien tavoitteena on tuottaa kotimainen näyttö siitä, miten semanttinen harmonisointi vaikuttaa tekoälymallien suorituskykyyn, datatuotannon ylläpidettävyyteen ja kokonaiskustannuksiin. Kansainvälinen näyttö on lupaavaa, mutta suomalainen referenssi puuttuu.

Ensimmäisenä askeleena julkaistaan pilottihaku vuoden 2026 kolmannella neljänneksellä ja valitaan toteuttajat vuoden loppuun mennessä. Pilottien kesto on 18 kuukautta, ja tulokset raportoidaan julki.

Johtopäätökset ja seuraukset

Terveystieteiden tekoälykeskustelussa nähtävä toistuva virhe on aloittaa arviointi tekoälymallista ja edetä taaksepäin dataan vasta kun tekoälymalli ei toimi. Oikea järjestys on päinvastainen: ensin rakennetaan semanttisesti yhtenäinen, versioitu, testattu ja omistettu tietopohja — sitten vasta koulutetaan luotettava malli.

Data as Software ei ole yksittäinen tuote tai alusta. Se on menetelmällinen siirtymä, jossa datantuotantoa johdetaan insinööriyönä ja semanttinen harmonisointilogiikka rakennetaan arkkitehtuuriin sisään. S2-kerros on se kohta arkkitehtuurissa, jossa heterogeeninen terveydenhuollon koodistoviidakko muuttuu kliinisesti mielekkäiksi, tekoälylle syötettäväksi käsitteyksiköiksi. Ilman tätä kerrosta tekoälyhankkeet kohtaavat väistämättä saman peruskriisin: malli on hyvä, mutta data ei kerro samaa tarinaa järjestelmästä toiseen.

Tuotantokäytön ja pilottikäytön ero on sama kuin ohjelmistokehityksessä: pilotti voi toimia laboratorio-olosuhteissa, mutta tuotantokäyttö edellyttää hallittuja syötteitä, julkaisulogiikkaa, drift-seurainta, eksplisiittistä omistajuutta ja ylläpidettävää elinkaarimallia. DaS-menetelmä rakentaa nämä edellytykset tietopohjan tasolle ennen kuin tekoälymalli kohtaa operatiivisen ympäristön haasteet.

Suomella on ainutlaatuinen etu: kansalliset terveysrekisterit, -datainfrastruktura ja yhtenäinen sosiaali- ja terveydenhuollon rakenne. Tämä etu realisoituu vain, jos tietopohjan semanttinen laatu on varmistettu arkkitehtuurissa — ei projektikohtaisesti, vaan rakenteellisesti ja pysyvästi.

Ville Siltala

Ville Siltalalla on yli 10 vuoden kokemus ohjelmistokehityksestä sekä yli 5 vuoden kokemus tietotalustoista ja datainsinööriyöstä. Hän on toiminut johtavan datainsinöörin ja data-arkkitehdin roolissa useissa terveydenhuollon asiakas- ja tuotekehitysprojekteissa, painopisteenään hyvinvointialueiden tietotalustaympäristöjen kehittäminen suunnittelusta tuotantoon. Hänellä on vahva kokemus migraatioprojekteista, joissa olemassa olevia datajärjestelmiä modernisoidaan lakehouse-teknologioilla. Ville Siltala vastaa Inviniten ohjelmisto- ja data-arkkitehtuurikehittämisestä erityisesti sote-alan tietotalustaratkaisujen ja datainsinööriyön alueella.

Severi Niskanen

Severi Niskasella on yli 15 vuoden kokemus datatyöstä sekä yli 8 vuoden kokemus datakonsultoinnista sosiaali- ja terveydenhuollon toimijoille. Hän on toteuttanut yli 100 asiakastoimeksiantoa tiedolla johtamisen, data-analytiikan ja tutkimuksen alueilla sekä toiminut neljä vuotta THL:n vierailevana tutkijana. Hänellä on syvä ymmärrys sote-sektorin tietovirroista, tietovarannoista ja viitearkkitehtuureista sekä käytännön kokemus dataputkien ja BI-ratkaisujen toteutuksesta. Severi Niskanen vastaa Inviniten analytiikkakehittämisestä sote-alan tiedolla johtamisen ja analytiikkaratkaisujen alueella.

Antti Brunni

Antti Brunni on kokenut strategi ja johtaja. Hänellä on noin 20 vuoden kokemus digitaalisen transformaation, luovan teknologian ja terveydenhuollon liiketoiminnan kehittämisestä. Taustaltaan Brunni on muotoilija ja liiketoimintastrategi; hän valmistelee väitöskirjaa lääketieteellisten ohjelmistojen suunnittelun strategioista. Antti Brunni toimii Inviniten toimitusjohtajana.

Aritikkeli on osa Inviniten Insights-julkaisusarjaa, jossa julkaistaan Inviniten asiantuntijoiden kirjoituksia ajankohtaisista aiheista sosiaali- ja terveydenhuollon digitalisaation alueella.



www.invinite.fi