Leevi Niinimäki

Master datan eheyttäminen ja rikastaminen tekoälytyökalujen avulla

Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Automaatio ja tietotekniikka
Diplomityö

Vaasa 2026

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tekijä: Leevi Niinimäki
Tutkielman nimi: Master datan eheyttäminen ja rikastaminen tekoälytyökalujen avulla
Tutkinto: Tietojenkäsittelytieteiden maisteriohjelma
Opintosuunta: Automaatio ja tietotekniikka
Työn ohjaajat: Timo Mantere
Janne Koljonen
Valmistumisvuosi: 2026 **Sivumäärä:** 79 (Liitteen kanssa 80)

TIIVISTELMÄ:

Teollisuusyritysten master data eli perustiedot ovat tärkeä osa yritysten toimintaa. Ne muodostavat perustan hankinnalle, myynnille sekä materiaalien ja tuoterakenteiden hallinnalle. Tämän takia master datan laatu on kriittinen tekijä prosessien sujumisessa ja onnistumisessa. Master datan heikko laatu vaikuttaa negatiivisesti prosessien tehokkuuteen, lisää virheiden riskiä ja manuaalisen työn määrää sekä rajoittaa mahdollisuutta prosessien automaatiolle. Tässä diplomityössä tarkastellaan projektia, jossa kehitetään tekoälypohjaista master datan hallinta- ja rikastustyökalua. Tutkimuksen kohdeyritys on Arnon Oy ja projekti toteutetaan osana yrityksessä käynnistettyä tekoälyhanketta.

Tutkimuksen tavoitteena oli analysoida, miten tekoälyä voidaan hyödyntää master datan eheyttämisessä, rikastamisessa ja hallinnassa. Lisäksi tavoitteena oli tunnistaa dataan, tekniseen toteutukseen ja organisaatioon liittyviä tekijöitä, jotka vaikuttavat tekoälypohjaisen työkalun onnistuneeseen käyttöönottoon. Tutkimus pohjautuu Design Science Research -lähestymistapaan, jossa projektia verrataan teknisen artefaktin kehitykseen. Työn empiirinen aineisto koostui nykytila-analyysista ja käyttäjäkyselystä. Nykytila-analyysissä master datalle tehtiin kokonaisvaltainen data-analyysi, jossa tutkittiin master datan laatua kuuden ulottuvuuden pohjalta. Kyselytutkimus suunnattiin master datan käyttäjille, joiden työhön kehitettävä työkalu vaikuttaa. Lisäksi työssä suunniteltiin ja analysoitiin tekoälytyökalun arkkitehtuuria, joka perustui Azure AI Foundryssa kehitettyyn Retrieval-Augmented Generation (RAG) malliin.

Tutkimuksen tulokset osoittivat, että kohdeyrityksen master data on rakenteeltaan kattava, mutta laadullisesti puutteellinen ja epätasainen. Keskeisimmät ongelmat master datassa olivat puutteelliset attribuutit, epäyhtenäiset nimeämiskäytännöt ja duplikaattinimikkeet. Nämä master datan laadulliset ongelmat heikentävät sekä datan käytettävyyttä että tekoälytyökalun toimintaa. Kyselytutkimuksen vastausten perusteella nykyiset manuaaliset prosessit koetaan työläiksi ja tekoälyavusteiseen datan eheyttämiseen, rikastamiseen ja hallintaan suhtaudutaan myönteisesti. Käyttäjät korostivat kuitenkin ihmisen tekemää lopullista tarkastusta, käyttäjän kontrollia, lähteiden käytön läpinäkyvyyttä ja vaiheittaisen käyttöönoton merkitystä.

Projektissa kehitetyllä tekoälytyökalulla datan eheytys toimii siten, että työkalu pystyy tunnistamaan duplikaatteja, epäyhtenäisiä nimeämisiä ja puutteellisia attribuutteja. Tämän jälkeen työkalu muodostaa korjausehdotuksia havaituista ongelmista ja antaa ne loppukäyttäjän tarkistettavaksi. Datan rikastamisessa työkalu tunnistaa puuttuvia tietoja ja hakee niitä ennalta määritellyistä lähteistä. Haettujen tietojen perusteella muodostetaan rikastusehdotus, jonka loppukäyttäjä voi tarkistaa ja hyväksyä. Master datan hallinta toimii siten, että käyttäjä voi kysyä

työkalua listaamaan tai etsimään nimikkeitä tiettyjen attribuuttien perusteella, joita ovat esimerkiksi mitat, kategoria tai materiaali.

Tutkimuksen perusteella tekoälypohjainen rikastustyökalu on toteuttamiskelpoinen kohdeyrityksessä, mutta vaatii vielä mallin kehitystä sekä master datan laadun korjausta. Projektissa kehitetty työkalu ei kyennyt riittävään tarkkuuteen viimeisessä kontrollissa, joten nykytilanteessa sitä ei voida ottaa tuotantokäyttöön. Työkalun jatkokehitys on merkittävä potentiaalinen kehityssuunta, mutta se vaatii samanaikaista master datan harmonisointia sekä panostusta tekniseen arkkitehtuuriin ja käyttäjälähtöiseen käyttöönottoon. Tämä työ tarjoaa käytännön suunnitelman tekoälytyökalun jatkokehitykselle ja tuottaa käytännönläheistä tietoa tekoälyn hyödyntämisestä master datan hallinnassa teollisuusyrityksen kontekstissa.

AVAINSANAT: Master data, Datan laatu, Datan eheyttäminen, Datan rikastaminen, Tekoäly, Retrieval-Augmented Generation (RAG), Design Science Research (DSR), Datanhallinta, tekoälytyökalu

Sisällys

Kuvat

Lyhenteet

AI	=	Artificial Intelligence (tekoäly)
BOM	=	Bill Of Materials (materiaaliluettelo)
DQM	=	Data Quality Management (datan laadunhallinta)
DSR	=	Design Science Research (design-tieteellinen tutkimus)
ERP	=	Enterprise Resource Planning (toiminnanohjausjärjestelmä)
MDG	=	Master Data Governance (master datan hallinnointi)
MDM	=	Master Data Management (master datan hallinta)
MES	=	Manufacturing Execution System (tuotannonohjausjärjestelmä)
MVP	=	Minimum Viable Product (minimivaatimukset täyttävä tuote)
NLP	=	Natural Language Processing (luonnollisen kielen käsittely)
PoC	=	Proof of Concept (todiste konseptin toimivuudesta)
RRF	=	Reciprocal Rank Fusion (hakutulosten yhdistämismenetelmä)

1 Johdanto

Tekoälysovellukset ovat nousseet keskeiseen rooliin teollisuuden digitalisoitumisessa, sillä ne mahdollistavat suurten ja monimutkaisten tietomassojen käsittelyn, prosessien automatisoinnin sekä reaaliaikaisen päätöksenteon. Koneoppimiseen ja tekoälyyn perustuvat työkalut tarjoavat uusia tapoja optimoida tuotantoa, parantaa laadunvalvontaa ja kehittää toimitusketjujen hallintaa (Culot ja muut, 2024). Tekoälyä hyödynnetään teollisuudessa esimerkiksi laitteiden ennakoivassa kunnossapidossa, älykkäissä valmistusjärjestelmissä ja datavetoisessa tuotesuunnittelussa, mikä vahvistaa yritysten kilpailukykyä globaalissa toimintaympäristössä (Dwivedi ja muut, 2023). Kansainväliset selvitykset ennustavat tekoälyn tuovan teollisuudelle huomattavia tuottavuushyötyjä ja McKinsey & Company (2023) arvioi sen lisäävän globaalin talouden arvoa useilla miljardeilla dollareilla vuosittain.

1.1 Taustakuvaus ja ongelmanasettelu

Arnon Oy käynnisti tekoälyhankkeen, jonka tarkoituksena on korvata manuaaliset ja virhealttiit prosessit automatisoiduilla AI-työkaluilla. Hankkeen tavoitteena on tunnistaa ja poistaa virheellinen data sekä yhtenäistää ja rikastaa sitä ulkopuolisella datalla. Datan analysoinnin ja eheyttämisen jälkeen on tarkoitus kehittää tekoälypohjaiset SAP-rikastustyökalu ja BOM-formaattityökalu, jotka integroidaan osaksi Arnonin prosesseja.

Nykytilassa Arnonin komponenttitietokannan *master datassa* on merkittäviä puutteita, mikä heikentää datan manuaalisen käytön tehokkuutta sekä uusien tekoälytyökalujen toimintaa. Komponenttitietokannan master data kärsii attribuuttien puuttumisesta, toimittajatietojen hajanaisuudesta sekä ristiriitaisista kuvauksista ja luokituksista. Nämä puutteet johtavat ylimääräisiin selvityksiin, manuaalisiin korjauksiin ja toimitusviiveisiin, heikentäen samalla sekä päivittäisten operaatioiden sujuvuutta että tekoälytyökalujen toimivuutta.

Tämän diplomityön tavoitteena on selvittää, miten master datan eheyttä ja laatua voidaan parantaa tekoälypohjaisen SAP-rikastustyökalun avulla ja miten datan siivous vaikuttaa Arnonin prosesseihin. Työssä tarkastellaan, millä menetelmillä master dataa voidaan eheyttää ja rikastaa, sekä arvioidaan, miten korjattu data tukee tulevia tekoälyhankkeita. Erityisesti tavoitteena on tuottaa käytännönläheistä tietoa data-analyysistä, rikastustyökalun käyttöönotosta ja haasteista, joita tekoälyprojekteissa voidaan kohdata. Tämän työn löydöksiä voidaan hyödyntää esimerkiksi projektin jatkokehityksessä ja tulevissa tekoälyprojekteissa. Tämä työ on toteutettu Teknologiateollisuuden 100-vuotissäätiön myöntämällä apurahalla.

1.2 Tutkimuksen tavoitteet

Tämän tutkimuksen yleisenä tavoitteena on analysoida Arnon Oy:n tekoälyhanketta, joka keskittyy komponenttitietokannan master datan parantamiseen ja tekoälyavusteisen rikastusprosessin kehittämiseen. Tutkimuksessa tarkastellaan erityisesti datan eheyttämistä ja täydentämistä sekä SAP-rikastustyökalun tehokkuutta ja käytettävyyttä osana yrityksen tiedonhallintaa.

1.2.1 Tutkimuksen päätavoite

Tutkimuksen päätavoitteena on selvittää, miten Arnonin master datan laatua voidaan parantaa tekoälyn ja automaation avulla siten, että ratkaisut tukevat organisaation liiketoimintaprosesseja ja datan hyödyntämistä laajemmassa mittakaavassa. Erityistä huomiota kiinnitetään datan eheyttämiseen, rikastamiseen ja validointiin osana SAP-rikastustyökalun kehitystyötä. Lisäksi arvioidaan nykyisen datan laatua, tunnistetaan keskeisimmät puutteet ja suunnitellaan menetelmiä, joilla näitä puutteita voidaan korjata.

Master datan laatu muodostaa perustan onnistuneelle tekoälyratkaisulle, sillä virheellinen tai epäyhtenäinen data johtaa heikkoihin algoritmien tuloksiin ja heikentää automaation tuottamaa arvoa. Master datan laatu on keskeinen edellytys onnistuneelle

tekoälyratkaisulle, sillä epäluotettava data johtaa heikkoihin algoritmien tuloksiin ja voi merkittävästi vähentää automaation tuottamaa hyötyä (Vallepu, 2025).

1.2.2 Operatiiviset tavoitteet

1. **Nykytila-analyysin suorittaminen.** Ensimmäisenä tavoitteena on kartoittaa nykyisen master datan laatu määrittämällä keskeiset laatumittarit, kuten täydellisyys, yhdenmukaisuus, saavutettavuus, tarkkuus ja ajantasaisuus (Batini & Scannapieco, 2016). Nykytila-analyysi antaa perustan myöhemmille kehitystoimille, sillä se auttaa arvioimaan ongelmien laajuutta ja tunnistamaan kriittisimmät pullonkaulat.
2. **Osallistuminen SAP-rikastustyökalun Proof-of-Concept (PoC) -vaiheeseen.** Tavoitteena on testata työkalua valitulla, rajatulla datajoukolla ja arvioida sen kykyä tunnistaa, korjata ja täydentää master dataa. PoC-vaihe mahdollistaa työkalun käytännön toimivuuden arvioinnin ennen laajamittaista käyttöönottoa, mikä on tärkeää riskienhallinnan ja resurssien kohdentamisen näkökulmasta (Peffers ja muut, 2007).
3. **Datan eheyttämis- ja rikastusprosessin analysointi ja testaaminen.** Tavoitteena on dokumentoida käytännön vaiheet, joilla puutteellista dataa parannetaan. Prosessien systemaattinen analyysi mahdollistaa niiden myöhemmän toistettavuuden sekä tarjoaa käytännönläheistä tietoa prosessien onnistumisista ja haasteista.
4. **Käytännön ohjeistuksen ja suositusten laatiminen.** Tutkimuksessa dokumentoidaan parhaat projektissa käytetyt menetelmät master datan hallintaan ja tekoälytyökalun käyttöönottoon. Näitä ohjeistuksia voidaan hyödyntää sekä hankkeen myöhemmissä vaiheissa että tulevissa tekoälyprojekteissa, mikä tukee tiedonhallinnan jatkuvaa kehittämistä organisaatiossa.
5. **Jatkokehitystä tukevan etenemismallin kehittäminen.** Viimeisenä tavoitteena on luoda projektin jatkokehitykselle suunnitelma ja käytännön suositukset

organisaatiolle. Tähän sisältyy myös onnistumisien ja haasteiden analysointi, mikä edesauttaa jatkokehitystä sekä tulevia projekteja.

1.3 Tutkimuskysymykset

Tämä tutkimus perustuu master datan eheyttämiseen ja rikastamiseen tekoälyn avulla. Master datan laatu ja sen analysointi on täten suuressa roolissa tässä tutkimuksessa, koska datan laadulla on suuri merkitys automaation onnistumiseen. Aiemmat tutkimukset ovat osoittaneet, että datan puutteet, epäyhtenäisyydet ja virheet heikentävät merkittävästi prosessien tehokkuutta ja automatisointimahdollisuuksia (McKinsey & Company, 2024). Tässä diplomityössä tutkimuskysymykset on muotoiltu seuraavasti:

1. **Mitkä master datan laatuun liittyvät attribuutit ovat kriittisimpiä tilausprosessin onnistumiselle?**
   Master datan laadun keskeiset ulottuvuudet, kuten täydellisyys, ajantasaisuus, yhdenmukaisuus ja tarkkuus, ovat osoittautuneet kriittisiksi tekijöiksi organisaatioiden operatiivisessa tehokkuudessa. Gualo ym. (2023) mukaan, nämä attribuutit vaikuttavat suoraan liiketoimintaprosessien virheettömyyteen, läpimenoaikoihin ja automaation luotettavuuteen erityisesti tilaus- ja toimitusprosesseissa. Lisäksi laadukas master data vähentää virheellisten tilausten määrää ja parantaa toimitusvarmuutta.
2. **Millä menetelmillä master dataa voidaan eheyttää ja rikastaa automaattisesti, ja mikä on niiden vaikutus datan laatuun?**
   Master datan automaattinen eheyttäminen ja rikastaminen hyödyntää yhä useammin tekoälyä ja koneoppimista. Näiden avulla voidaan tunnistaa virheitä, täydentää puuttuvia tietoja ja varmistaa datan yhdenmukaisuus ilman manuaalista työtä. Esimerkiksi generatiivinen tekoäly voi tuottaa puuttuvaa kontekstitietoa, kun taas älykkäät hakujärjestelmät ja validointimallit tukevat datan ajantasaisuutta ja tarkkuutta. Tutkimusten mukaan automaattiset

menetelmät parantavat merkittävästi datan laatua ja nopeuttavat sen ylläpitoa (Dwivedi ja muut, 2023).

3. **Kuinka kaupallinen AI-rikastustyökalu suoriutuu verrattuna manuaalisiin menetelmiin master datan rikastuksessa?**
   Tässä tutkimuksessa selvitetään miten tekoälyavusteinen master datan rikastaminen vertautuu manuaaliseen rikastamiseen. Lisäksi tutkitaan, saadaanko tekoälytyökalusta merkityksellistä hyötyä organisaation liiketoimintaan.
4. **Miten tekoälypohjaiset ratkaisut voivat täydentää perinteistä rikastustyötä?**
   Tekoälypohjaiset lähestymistavat täydentävät perinteisiä datan rikastusmenetelmiä tunnistamalla piileviä yhteyksiä ja virheitä, joita manuaaliset tarkistusprosessit eivät havaitse. Tekoäly voi automatisoida datan profiloinnin, poikkeamien tunnistamisen ja duplikaattien poiston, mikä vähentää virheitä ja parantaa päätöksenteon tarkkuutta (Ehrlinger ja muut, 2022). Lisäksi generatiiviset mallit voivat tuottaa puuttuvaa tai epävarmaa tietoa tukevia ennusteita, jolloin datan eheys vahvistuu.
5. **Mitä muutoksia tarvitaan, jotta eheä ja rikastettu master data voidaan ylläpitää pysyvästi?**
   Master datan eheyttäminen ja rikastaminen ei ole kertaluonteinen tavoite, vaan jatkuvaa hallintaa vaativa prosessi. Ibrahim ym. (2021) korostavat, että pysyvän laadun ylläpito edellyttää systemaattista hallintamallia, jossa yhdistyvät teknologiset ratkaisut, organisatorinen vastuunjako ja säännöllinen laadunvalvonta. Tutkimus painottaa erityisesti jatkuvaa arviointia, versionhallintaa ja käyttäjien osallistamista laadun ylläpitoon.

1.4 Tutkielman rakenne

Tämä tutkielma rakentuu kuudesta pääluvusta, jotka etenevät tutkimusongelman pohjustuksesta empiiriseen analyysiin ja johtopäätöksiin.

Luvussa 1 esitellään tutkimuksen tausta, tavoitteet ja tutkimuskysymykset sekä perustellaan aiheen merkitys. Lisäksi kuvataan tutkimuksen rajaukset ja kokonaisrakenne.

Luvussa 2 esitetään kirjallisuuskatsaus ja teoreettinen viitekehys, jossa tarkastellaan master dataa, datan laatua sekä datan eheytykseen ja rikastamiseen liittyviä käsitteitä. Lisäksi luvussa käsitellään tekoälyn roolia datan laadun parantamisessa sekä SAP- ja BOM-prosesseja tutkimuksen kontekstissa. Luvun lopussa muodostetaan synteesi keskeisistä teoreettisista näkökulmista, jotka ohjaavat tutkimuksen toteutusta.

Luvussa 3 kuvataan tutkimusmenetelmät. Luvussa esitellään tutkimusstrategiana käytetty Design Science Research (DSR) -lähestymistapa sekä aineistonkeruumenetelmät, joihin kuuluvat nykytila-analyysi ja kyselytutkimus. Lisäksi kuvataan aineiston käsittely ja analysointi sekä tutkimuksen toteutuksen keskeiset vaiheet.

Luvussa 4 esitetään tutkimuksen tulokset ja analyysi. Tässä luvussa kuvataan kehitetyn tekoälyratkaisun tekninen toteutus ja arkkitehtuuri, analysoidaan master datan nykytila datan laadun näkökulmasta sekä tarkastellaan kyselytutkimuksen tuloksia käyttäjälähtöisestä näkökulmasta. Luvun lopussa esitetään tulosten synteesi sekä skenaariopohjainen analyysi projektin jatkokehityksestä.

Luvussa 5 käsitellään tulosten pohdinta, arvioidaan projektin teknisiä ratkaisuja, tarkastellaan tekoälyn vaikutuksia asiantuntijatyöhön sekä analysoidaan tutkimuksen luotettavuutta ja rajoitteita.

Luvussa 6 esitetään tutkimuksen johtopäätökset, vastataan tutkimuskysymyksiin, esitetään keskeiset havainnot sekä annetaan suosituksia kohdeyritykselle. Lisäksi esitetään mahdollisia jatkotutkimusaiheita ja jatkokehityksen suuntaviivoja.

2 Kirjallisuuskatsaus ja teoreettinen viitekehys

Master datan laatu muodostaa keskeisen lähtökohdan tämän diplomityön tarkastelulle, sillä sen eheys ja johdonmukaisuus vaikuttavat suoraan teollisten prosessien luotettavuuteen ja tehokkuuteen. Kirjallisuuskatsauksessa tarkastellaan aiempaa tutkimusta master datasta ja sen hallinnasta, datan laadun eri ulottuvuuksista sekä niiden merkityksestä erityisesti teollisuuden toimintaympäristöissä. Teoreettisen viitekehyksen avulla jäsennetään niitä käsitteitä, periaatteita ja menetelmiä, joiden varaan tutkimuksen analyysi rakentuu. Viitekehyksen avulla muodostetaan kokonaiskuva siitä, miten master data liittyy teollisuuden digitalisaatioon, automaatioon ja tekoälyratkaisujen hyödyntämiseen. Sen avulla myös selvitetään millä tavoin master datan hallinta tukee yrityksen kilpailukykyä ja operatiivista tehokkuutta (McKinsey & Company, 2024).

2.1 Master data ja sen merkitys teollisuudessa

Teollisuuden digitalisoituminen ja globaalit toimitusketjut ovat lisänneet merkittävästi datan määrää ja monimutkaisuutta. Tämän kehityksen keskellä master data, eli organisaation keskeiset tietokannat (tuote-, asiakas-, toimittaja- ja materiaalitiedot), on noussut yhdeksi tärkeimmistä tuotantoon ja liiketoimintaan vaikuttavista tekijöistä (Pansara, 2023).

Master Data Management (MDM) tarkoittaa näiden perustietojen hallintaa, yhtenäistämistä ja ylläpitoa koko organisaation tasolla. Se toimii perustana muun muassa tuotesuunnittelulle, tuotanto-ohjeille, toimittajahallinnalle ja asiakastiedon hallinnalle. Ilman eheää ja ajantasaista master dataa tuotantoprosessit altistuvat virheille ja viivästyksille. Esimerkiksi virheelliset tuotetiedot voivat johtaa tuotantokatkoksiin tai takaisinvetotilanteisiin (Pansara, 2023).

Tutkimusten mukaan jopa 68 prosenttia toimitusketjujen tehottomuudesta voidaan jäljittää huonoon datan laatuun ja epäyhtenäiseen master dataan. MDM:n hyötyjä ovat

esimerkiksi lyhyemmät käsittelyajat, parempi varaston tarkkuus, nopeampi toimittajien käyttöönotto sekä merkittävä investointien takaisinmaksu. Lisäksi MDM mahdollistaa reaaliaikaisen päätöksenteon, koska tieto on keskitettyä, yhdenmukaista ja luotettavaa. Reaaliaikaista päätöksentekoa tukee myös helpottunut integraatio keskeisiin yritysjärjestelmiin, kuten toiminnanohjausjärjestelmä (ERP, *Enterprise Resource Planning*) ja tuotannonohjausjärjestelmä (MES, *Manufacturing Execution System*). (Nair, 2025)

2.2 Datan laatu

Master data muodostaa liiketoimintaprosessien perustan, ja sen eheys sekä laatu vaikuttavat suoraan päätöksenteon luotettavuuteen, tehokkuuteen ja sääntelyn noudattamiseen. Eheä data tarkoittaa, että tiedot ovat täsmällisiä, ajantasaisia, yhdenmukaisia ja liiketoimintasääntöjen mukaisia. Kun data on eheää, se tukee liiketoimintaprosesseja saumattomasti ja vähentää virheiden riskiä.

Esimerkiksi Vihavaisen (2014) tutkimus osoittaa, että eheä asiakasdata vähentää käyttäjien aikaa tiedon etsimiseen ja parantaa organisaation luotettavuutta.

Datan laatu ei ole pelkästään tekninen kysymys, vaan strateginen tekijä, joka vaikuttaa suoraan liiketoiminnan sujuvuuteen, päätöksenteon luotettavuuteen ja tekoälyratkaisujen onnistumiseen. Huonolaatuinen data, kuten päivittämättömät tiedot, duplikaatit ja epäyhtenäiset nimikkeet, aiheuttaa merkittäviä viiveitä, virheitä ja lisäkustannuksia organisaatioille. Esimerkiksi asiakkuudenhallintajärjestelmissä duplikaattitietojen osuus voi nousta jopa 20 prosenttiin, mikä johtaa asiakastietojen sekaannuksiin ja heikentää operatiivista tehokkuutta. (Kim, 2025)

Raportissa korostetaan, että datan huono laatu ei ole enää yksittäinen tai tilapäinen ongelma, vaan laaja-alainen kriisi, joka uhkaa tekoälyhankkeiden onnistumista ja liiketoiminnan tavoitteiden saavuttamista. Datan eheys ja ajantasaisuus ovat siten välttämättömiä edellytyksiä kaikille organisaatioille, jotka pyrkivät hyödyntämään automaatiota ja tekoälypohjaisia ratkaisuja tehokkaasti. (Kim, 2025)

Data Quality Management (DQM) tarkoittaa jatkuvaa prosessia, jonka tavoitteena on varmistaa, että data on tarkkaa, johdonmukaista, ajantasaista ja käyttötarkoitukseensa soveltuvaa. DQM kattaa kaikki vaiheet tiedon keräämisestä ja tallentamisesta sen validointiin, puhdistamiseen, rikastamiseen ja jatkuvaan valvontaan. Tehokas DQM-strategia varmistaa, että data on sekä teknisesti eheää että liiketoiminnan tarpeita vastaavaa, mikä tukee luotettavaa päätöksentekoa ja mahdollistaa prosessien automaation. (Sargiotis, 2024)

Keskeisiä datan laadun parantamisen käytäntöjä ovat tietojen puhdistaminen ja rikastaminen, yhtenäisten liiketoimintasääntöjen määrittely, selkeät roolit ja vastuut sekä tietojärjestelmien integrointi ja standardointi. Kun eheys ja laatu ovat kunnossa, organisaation päätökset perustuvat oikeaan tietoon ja prosessit toimivat tehokkaasti. (Ibrahim ja muut, 2021)

2.3 Master datan eheys ja rikastaminen

Master datan eheys ja rikastaminen liittyvät suurelta osin toisiinsa, mutta niiden merkitys eroaa huomattavasti. Eheä data tarkoittaa sitä, että tiedot ovat tarkkoja, yhdenmukaisia ja ajantasaisia. Datan eheyttä mitataan erilaisilla ulottuvuuksilla, jotka keskittyvät datan laadun eri osa-alueisiin (Batini & Scannapieco, 2016). Rikastaminen puolestaan tarkoittaa olemassa olevan datan täydentämistä tai standardointia, esimerkiksi lisäämällä puuttuvia attribuutteja, yhtenäistämällä nimeämisiä tai tuomalla mukaan ulkoisia tietolähteitä. Valmiiksi eheää master dataa on helpompi rikastaa uusilla tietolähteillä, koska taustadatan ollessa kunnossa, yhteydet ja merkitykset löytyvät datarakenteista helpommin ja tehokkaammin (Vihavainen, 2014).

Monet nykyiset ratkaisut hyödyntävät tekoälyä, ja koneoppiminen on erityisen tehokasta monistuneiden tai hieman erimuotoisten tietueiden tunnistamisessa ja yhdistämisessä. Perinteiset säännöt ja heuristiikat toimivat hyvin yksinkertaisissa tapauksissa, mutta

oppivat mallit pystyvät löytämään monimutkaisia ja epäsuoria yhteyksiä eri kenttien välillä, mikä parantaa tarkkuutta. (Rana ja muut, 2024)

Luonnollisen kielen käsittelyä käytetään rikastamisessa silloin, kun tieto on tekstiä tai vapaamuotoista kuvausta. Natural Language Processing (NLP) -menetelmät auttavat muuttamaan vapaata tekstiä rakenteelliseksi tiedoksi, jolloin esimerkiksi tuotetekstit, kommentit tai vapaamuotoiset kuvaukset voidaan linkittää master dataan. Tämä vähentää manuaalista työtä ja mahdollistaa laajemman automaation. (Lewis ja muut, 2020; Rana ja muut, 2024)

Tekoälyllä on myös kyky oppia organisaation omista säännöistä ja käytännöistä, joten datan laadun hallinta voi muuttua reaktiivisesta korjaamisesta jatkuvasti paranevaksi prosessiksi. Tällainen adaptiivinen hallinta sopeutuu paremmin suurten tietomassojen kanssa ja tukee aktiivista laadunvalvontaa. On kuitenkin tärkeää yhdistää teknologia selkeisiin prosesseihin ja vastuisiin, jotta ratkaisut toimivat luotettavasti käytännössä. (Lewis ja muut, 2020)

2.4 Datan profilointi ja laadun mittausmenetelmät

Datan profilointi ja laadun mittausmenetelmät ovat keskeisiä vaiheita datan hallintaprosessissa, erityisesti silloin kun pyritään varmistamaan tiedon eheys, käyttökelpoisuus ja luotettavuus. Profiloinnilla tarkoitetaan datan rakenteen, sisällön ja ominaisuuksien analysointia. Laadun mittausmenetelmien avulla arvioidaan, kuinka hyvin data täyttää sille asetetut laatuvaatimukset esimerkiksi täydellisyyden, tarkkuuden, ajantasaisuuden ja yhdenmukaisuuden osalta. (Ehrlinger & Wöß, 2022)

Datan profilointi on keskeinen lähtökohta datan laadunhallinnalle, sillä sen avulla voidaan arvioida datan laatua ennen jatkokäsittelyä (Ehrlinger & Wöß, 2022). Se auttaa tunnistamaan poikkeavuuksia, puuttuvia arvoja, duplikaatteja ja epäjohdonmukaisuuksia, jotka voivat heikentää datan käyttökelpoisuutta. Profiloinnin avulla voidaan myös määrittää datan jakautumista, arvojen yleisyyttä ja kenttien

riippuvuuksia, mikä tukee myöhempiä siivous- ja rikastusvaiheita. Artikkeli osoittaa, että monissa käytännön työkaluissa mittareiden hyödyntäminen on puutteellista, mikä korostaa tarvetta kehittää tehokkaampia ja käyttäjäystävällisempiä ratkaisuja.
Tutkimuksessa ”Artificial intelligence methods and approaches to improve data quality in healthcare data” todetaan, että tekoälyä voidaan hyödyntää merkittävästi muun muassa datan profiloinnissa sekä laadun mittaamisessa (Agate, 2025). Esimerkiksi koneoppimismallit voivat automaattisesti tunnistaa virheellisiä arvoja, puuttuvia tietoja ja epäloogisia riippuvuuksia datassa. Tämä mahdollistaa datan laadun jatkuvan seurannan ja parantamisen ilman manuaalista työtä.

2.5 Tekoälyn rooli datan laadun parantamisessa

Tekoälystä on tullut keskeinen työkalu datan laadunhallinnassa, se pystyy käsittelemään ja analysoimaan suuria tietomääriä huomattavasti tehokkaammin kuin manuaaliset tarkistukset tai perinteiset sääntöpohjaiset ratkaisut. Viimeaikaiset tutkimuskatsaukset osoittavat, että tekoälyä voidaan hyödyntää datan laadun valvonnassa koko prosessin laajuudelta: profiloinnissa, virheiden tunnistamisessa, sääntöjen muodostamisessa ja laadun jatkuvassa seurannassa. Vaikka tekoälyyn perustuvia sovelluksia on jo otettu käyttöön useilla aloilla, niiden potentiaali on edelleen osin hyödyntämättä (Tamm & Nikiforova 2024).

Koneoppimismallit auttavat tunnistamaan poikkeamia, yhdistämään päällekkäisiä tietueita ja ehdottamaan sääntöjä, joilla datan laatua voidaan parantaa. Näiden menetelmien avulla on mahdollista vähentää manuaalista työtä ja nopeuttaa korjauskierroksia erityisesti monilähteisissä ja teknisesti hajanaisissa järjestelmissä (Tamm & Nikiforova 2024).

Generatiivinen tekoäly laajentaa perinteisten menetelmien sovellusaluetta tilanteissa, joissa data on puutteellista tai osin tekstimuotoista. Mallit voivat esimerkiksi täydentää metatietoja, täyttää puuttuvia arvoja ja muuntaa tekstisisältöjä rakenteiseen muotoon,

mikä helpottaa rikastamista ja validointia. Näiden menetelmien vaikutuksia on tutkittu erityisesti datalaadun kannalta kriittisillä aloilla, kuten terveydenhuollossa (Agate, 2025).

Sekä tutkimus että käytännön kokemukset osoittavat, että tekoälyratkaisujen onnistuminen riippuu suoraan lähtöaineiston laadusta. Puutteellinen tai epäyhtenäinen data heikentää mallien tarkkuutta ja luotettavuutta, minkä vuoksi datan laadun parantaminen tulisi toteuttaa ennen tekoälymallien käyttöönottoa (Ribeiro 2024; Jarrahi ja muut, 2024).

2.5.1 Tekniset menetelmät

Tekoäly tarjoaa useita konkreettisia keinoja datan laadun parantamiseen. Ensinnäkin se mahdollistaa virheiden automaattisen tunnistamisen ja korjaamisen, kuten puuttuvien arvojen täyttämisen, typografisten virheiden korjaamisen ja ristiriitaisten tietueiden havaitsemisen. Tutkimukset osoittavat, että automaattiset korjausalgoritmit voivat merkittävästi lisätä datan tarkkuutta ja vähentää manuaalisen työn tarvetta (Chu ja muut, 2016; Rekatsinas ja muut, 2017).

Toiseksi tekoäly tukee metatietojen rikastamista ja standardointia. Mallit voivat generoida ja yhdenmukaistaa metatietoja, mikä parantaa tiedon löydettävyyttä, ymmärrettävyyttä ja integroitavuutta eri lähteistä. Tutkimukset osoittavat, että metadatan hallinta tekoälymenetelmin voi merkittävästi tehostaa järjestelmien suorituskykyä ja tiedonhallintaa (Yang, Fu & Amin, 2025).

Kolmanneksi tekoäly on tehokas väline poikkeavuuksien ja duplikaattien tunnistamisessa. Koneoppimismallit pystyvät käsittelemään suuria tietomääriä ja löytämään niistä päällekkäisyyksiä, poikkeavia arvoja ja tietueita, jotka eivät noudata ennalta määriteltyjä rakenteita. Datan yksilöllisyyttä ja tulkinnan tarkkuutta voidaan parantaa huomattavasti tekoälyyn perustuvalla kaksoiskappaleiden tunnistamisella (Mu ja muut, 2024).

Neljänneksi tekoälyllä voidaan tehostaa datan laadun jatkuvaa seurantaa ja raportointia. Datahavainnointijärjestelmät, jotka hyödyntävät tekoälyyn perustuvia profilointi- ja poikkeamahavaintoja, mahdollistavat datan tilan seurannan reaaliajassa ja tukevat nopeaa reagointia havaittuihin poikkeamiin (Ehrlinger ja muut, 2022).

Lopuksi tekoäly voi automatisoida sääntöjen generoinnin ja validoinnin. Sen sijaan, että laatua koskevia sääntöjä määriteltäisiin manuaalisesti, mallit voivat oppia uusia sääntöjä datasta tunnistettujen kuvioiden perusteella ja arvioida niiden tehokkuutta jatkuvasti. Tämä lähestymistapa tukee laadunhallintaa suurissa ja nopeasti muuttuvissa tietomassoissa (Tamm & Nikiforova, 2024).

2.5.2 Organisatoriset näkökulmat ja soveltamisen edellytykset

Vaikka tekoäly tarjoaa tehokkaita teknisiä ratkaisuja, niiden käytännön hyödyntäminen riippuu myös organisaation rakenteista, osaamisesta ja prosesseista. Tamm ja Nikiforova (2024) korostavat, että tekoälyyn perustuvien laatumenetelmien onnistunut käyttöönotto edellyttää selkeitä vastuita, laadunhallinnan prosesseja ja datan omistajuuden määrittelyä.

Lisäksi henkilöstön osaaminen ja luottamus tekoälyyn vaikuttavat merkittävästi sen käyttöön. Datan laadun parantaminen tekoälyn avulla onnistuu parhaiten, kun tekninen kehitys yhdistetään koulutukseen ja organisaation sisäiseen viestintään (Culot ja muut, 2024; Ribeiro, 2024).

Organisatorinen näkökulma täydentää teknistä analyysia osoittamalla, että datan laadun kehittäminen on ennen kaikkea muutosprosessi, joka edellyttää sekä teknologisia että inhimillisiä valmiuksia.

2.6 SAP ja BOM-prosessit datan hyödyntämisen kontekstina

SAP-järjestelmässä master data, kuten materiaalitiedot sekä tuoterakenteet (BOM, Bill of Materials), muodostavat perustan monille teollisuuden prosesseille. Tuoterakenteet määrittelevät, miten tuotteet koostuvat eri komponenteista ja materiaaleista, ja niiden virheet voivat aiheuttaa viiveitä hankinnassa, tuotannon keskeytyksiä sekä lisäkustannuksia. Siksi BOM-tiedon ajantasaisuus, tarkkuus ja täydellisyys ovat välttämättömiä tuotannon luotettavuuden ja tehokkuuden kannalta (SAP, 2025).

Master datan hallinta SAP-ympäristössä sisältää muun muassa materiaalien, asiakkaiden ja toimittajien tietojen ylläpidon. Näiden tietojen oikeellisuus vaikuttaa suoraan ERP-prosessien onnistumiseen ja koko toimitusketjun sujuvuuteen. Tutkimusten mukaan heikkolaatuinen data aiheuttaa merkittävän osan teollisuuden prosessien tehottomuudesta, ja jopa 40 prosenttia toimitusketjun ongelmista voidaan jäljittää virheelliseen tai puutteelliseen master dataan (Matayo, S., 2026).

SAP tarjoaa työkaluja, kuten SAP Master Data Governance (MDG), joiden avulla tietojen hallintaa voidaan tehostaa. Tässä diplomityössä keskitytään kuitenkin erilliseen SAP-rikastamistyökaluun, joka kehitetään parantamaan datan laatua ja yhtenäisyyttä SAP-järjestelmän sisällä. Työssä edetään suunnittelusta ja Proof of Concept (PoC) -vaiheesta kohti minimikelpoisen tuotteen (MVP) toteutusta ja käytön arviointia. Työkalun tarkoituksena on tunnistaa ja täydentää puuttuvia tietoja, korjata epäjohdonmukaisuuksia sekä varmistaa tietojen oikeellisuus ennen niiden hyödyntämistä tuotannon ja hankinnan prosesseissa (McKinsey & Company, 2024).

Tämä diplomityö luo samalla pohjan seuraavalle kehitysvaiheelle, jossa toinen opiskelija toteuttaa tekoälypohjaisen BOM-työkalun. Tämän jatkotyön tavoitteena on automatisoida tuoterakenteiden luonnin ja hallinnan prosesseja hyödyntämällä rikastettua ja luotettavaa dataa. Tässä työssä BOM-prosessia käsitellään vain rajatusti, mutta sen merkitys on tärkeä, sillä se kytkeytyy suoraan datan laatuun ja eheään master

dataan. Kun tiedot ovat tarkkoja ja yhdenmukaisia, ne mahdollistavat luotettavan pohjan myös tuleville tekoälyratkaisuille (SAP, 2025; McKinsey & Company, 2024).

2.7 Retrieval-Augmented Generation ja kielimallipohjaiset hakuratkaisut

Retrieval-Augmented Generation (RAG) on lähestymistapa, jossa generatiivisen kielimallin tuottamaa vastausta täydennetään ulkoisesta tietovarannosta haetulla kontekstilla. Menetelmä yhdistää parametristen mallien generatiivisen kyvykkyyden ja ei-parametrisen, haun avulla ylläpidettävän tietopohjan, mikä parantaa vastauksien faktuaalista tarkkuutta ja mahdollistaa tiedon päivittämisen ilman mallin uudelleenkoulutusta (Lewis ja muut, 2020). RAG-menetelmä soveltuu erityisen hyvin tilanteisiin, joissa vastauksilta edellytetään organisaatiokohtaista, ajantasaista ja lähteistettävää tietoa. Tässä tutkimuksessa RAG toimii master datan rikastamisen teknisenä perustana, koska sen avulla voidaan yhdistää sisäinen nimikekirjasto, toimittajadata ja muut hyväksytyt tietolähteet yhtenäiseksi hakukontekstiksi. Menetelmä tukee siten sekä datan eheyden parantamista että käyttäjälle näkyvää läpinäkyvyyttä (Microsoft, 2026).

2.8 Yhteenveto teoreettisesta viitekehyksestä

Kirjallisuuskatsauksen perusteella voidaan todeta, että master datan laatu ja hallinta ovat keskeisiä tekijöitä teollisuuden tietojärjestelmien toimivuuden ja luotettavuuden kannalta. Hyvälaatuinen data tukee päätöksentekoa, lyhentää käsittelyaikoja ja vähentää virheitä tuotannossa ja toimitusketjuissa. Sen sijaan puutteellinen data voi johtaa merkittäviin taloudellisiin menetyksiin ja toimintakatkoksiin. Datan laatuun liittyvät ongelmat eivät ole pelkästään teknisiä, vaan ne vaikuttavat suoraan yritysten liiketoimintatavoitteiden saavuttamiseen. Aiemmat tutkimukset osoittavat, että laadukas master data voi tuoda jopa yli kaksinkertaisen investoinnin takaisinmaksun (Return Of Investment, ROI) kolmen vuoden sisällä (Ibrahim ja muut, 2021; Kim, F., 2025).

Teoreettinen viitekehys korostaa myös Data Quality Managementin (DQM) ja tekoälyn yhteyttä. DQM tarjoaa rakenteen datan hallintaan, jossa korostuvat tiedon tarkkuus, ajantasaisuus ja käytettävyyden jatkuva seuranta. Tekoäly puolestaan tuo uusia mahdollisuuksia näiden tavoitteiden saavuttamiseen, sillä se kykenee tunnistamaan virheitä, täydentämään puuttuvia tietoja ja oppimaan datan taustalla olevia riippuvuuksia. Näin tekoäly voi tukea jatkuvaa laadunparannusta ja vähentää manuaalista työtä.

On kuitenkin huomioitava, että tekoälyratkaisujen luotettavuus on suoraan riippuvainen käytetyn datan laadusta. Jos lähtödata sisältää virheitä tai epäjohdonmukaisuuksia, myös tekoälymallien tuottamat tulokset voivat olla epätarkkoja. Tämän vuoksi datan laadun parantaminen on tärkeää ennen tekoälytyökalujen käyttöönottoa. Synteettinen data tarjoaisi mahdollisuuden testata tekoälyratkaisuja ilman arkaluontoisia tietoja, mutta sen käyttö edellyttää tarkkaa arviointia. Tässä projektissa synteettinen data ei ole tarpeellista, koska komponenttitietokannan tiedot eivät arkaluontoisia ja testaus voidaan suorittaa rajatulla joukolla.

Tässä diplomityössä muodostettu teoreettinen viitekehys tarjoaa perustan SAP-rikastamistyökalun suunnittelulle ja arvioinnille. Sen avulla voidaan ymmärtää, miksi datan eheys ja laatu ovat välttämättömiä SAP-järjestelmien tehokkaalle toiminnalle ja miten tekoälyä voidaan hyödyntää datan rikastamisessa. Tämä kokonaisuus antaa myös pohjan myöhemmälle BOM-työkalun kehitystyölle, joka hyödyntää tämän työn tuloksena syntyvää parannettua dataa. Näin kirjallisuuskatsaus yhdistää teoreettisen tiedon ja käytännön sovelluksen, jotka yhdessä tukevat yrityksen pyrkimystä datalähtöiseen päätöksentekoon ja automaatioon.

3 Tutkimusmenetelmät

Tämän luvun tarkoituksena on kuvata tutkimuksessa käytetyt menetelmät, niiden valintaperusteet ja toteutustapa Arnon Oy:n tekoälyhankkeessa. Tutkimusmenetelmien avulla pyritään muodostamaan sekä laadullisesti että määrällisesti perusteltu kokonaiskuva siitä, miten datan eheyttä ja laatua voidaan parantaa tekoälyn ja automaation avulla osana teollisen yrityksen tietojärjestelmiä.

Tutkimus toteutettiin tapaustutkimuksena, jossa yhdistettiin useita eri menetelmiä: data-analyysiä, haastatteluja, kyselyaineistoa ja käytännön validointia datan rikastamisen prosessissa. Näiden lähestymistapojen yhdistäminen mahdollisti sekä teknisen että organisatorisen näkökulman huomioimisen, mikä on olennaista datan eheytys- ja rikastusratkaisujen onnistuneessa käyttöönotossa. (Bell, 2023)

Ensimmäisessä luvussa esitellään tutkimusstrategia ja projekti Design Science Research-näkökulmasta. Toisessa luvussa käydään läpi aineistonkeruumenetelmät, kohdeyrityksen dataympäristö ja kehityshankkeen tavoitteet. Kolmannessa luvussa kuvataan nykytila-analyysin tarkoitus ja eteneminen.

Neljäs luku käsittelee henkilöstölle toteutettua kyselyä, jonka avulla selvitettiin käyttäjien kokemuksia, näkemyksiä ja tarpeita datan laadun ja prosessien kehittämiseksi. Tämä aineisto tuo tutkimukseen inhimillisen ja prosessilähtöisen näkökulman teknisen tarkastelun rinnalle.

Kokonaisuutena tämän luvun tavoitteena on tarjota selkeä ja järjestelmällinen kuvaus tutkimusprosessista siten, että menetelmien valinta ja soveltaminen ovat perusteltuja ja toistettavissa. Menetelmävalinnoissa on painotettu sekä teknistä luotettavuutta että käytännön hyödynnettävyyttä. Tekninen luotettavuus viittaa data-analyysin oikeellisuuteen sekä prosessien seurattavuuteen ja käytännön hyödynnettävyys viittaa tutkimuksen soveltuvuuteen yrityksen olemassa olevien järjestelmien kanssa. Tällä

tavoitellaan sitä, että tutkimuksen tulokset eivät jää vain teoreettiselle tasolle, vaan tukevat käytännössä kohdeyrityksen datanhallinnan ja tekoälykehityksen jatkovaiheita.

3.1 Tutkimusstrategia

Tämän tutkimuksen metodologisena viitekehyksenä hyödynnetään Design Science Research (DSR) -lähestymistapaa, joka on vakiintunut tutkimusstrategia tietojärjestelmätieteissä. Hevnerin ym. (2004) mukaan DSR-tutkimuksen ytimenä on innovatiivisten artefaktien kehittäminen ja arviointi monimutkaisten liiketoimintaongelmien ratkaisemiseksi. Tällaisia artefakteja ovat esimerkiksi algoritmit, mallit, käyttöliittymät tai kokonaisvaltaiset ohjelmistoratkaisut. Toisin kuin puhtaasti kuvaileva tai selittävä tutkimus, DSR on luonteeltaan pragmaattista ja tavoitteellista, joten siinä pyritään luomaan uutta tietoa suunnittelemalla jotain, mikä parantaa nykytilaa.

Tutkimusprosessi noudattaa iteratiivista sykliä, joka voidaan jakaa neljään keskeiseen vaiheeseen:

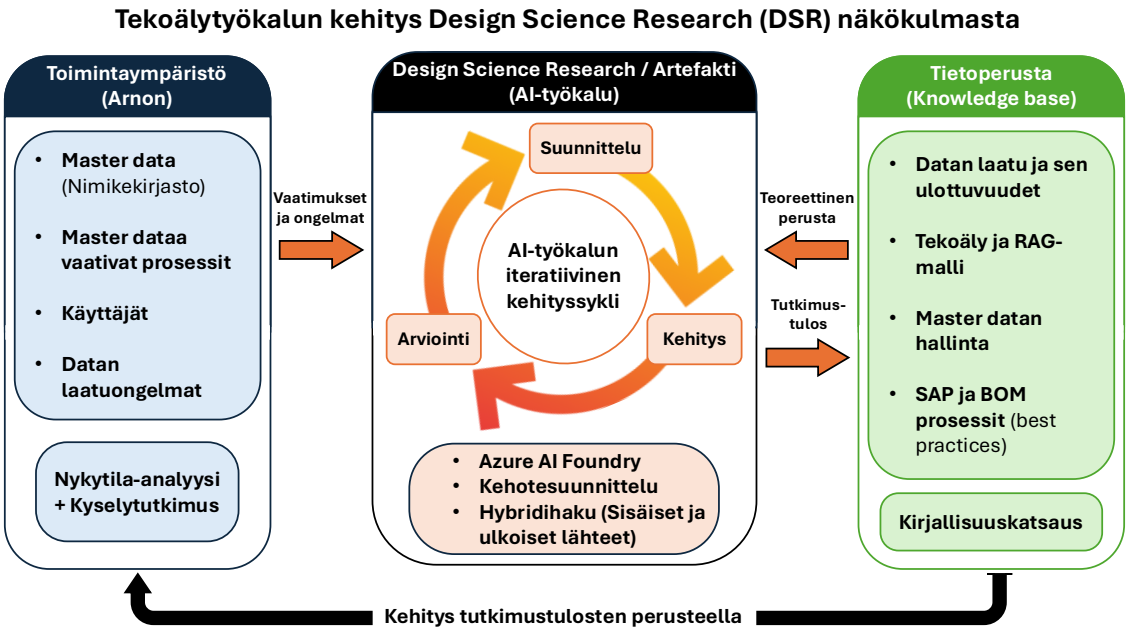
1. Ymmärtämisvaiheessa tunnistetaan kohdeyrityksen liiketoimintaympäristön haasteet ja määritetään tutkimusongelma. Tässä työssä vaihe kattaa master datan nykytila-analyysin ja käyttäjien tarpeiden kartoituksen.

2. Suunnitteluvaiheessa muodostetaan ratkaisun konseptuaalinen arkkitehtuuri hyödyntäen teoreettista tietoperustaa, kuten kirjallisuuskatsauksessa esiteltyjä tekoälymenetelmiä ja datan laadun viitekehyksiä.

3. Kehitysvaiheessa suunnitelma muunnetaan tekniseksi artefaktiksi. Tässä tutkimuksessa kehitysvaiheen tuloksena syntyy Azure-ympäristöön rakennettu, RAG-arkkitehtuuria hyödyntävä SAP-rikastustyökalu.

4. Arviointivaiheessa artefaktin suorituskykyä ja hyödyllisyyttä testataan suhteessa asetettuihin tavoitteisiin. Arviointi tuottaa palautetta, joka ohjaa joko artefaktin jatkokehitystä tai vahvistaa sen soveltuvuuden aiottuun käyttötarkoitukseen.

Tämä lähestymistapa varmistaa, että kehitetty tekoälyratkaisu ei ole vain tekninen kokeilu, vaan se on teoreettisesti perusteltu ja vastaa suoraan kohdeorganisaation käytännön ongelmaan. Projektin DSR-sykli on esitettynä kuvassa 1 (Kuva 1).



Kuva 1. AI-pohjaisen master datan rikastustyökalun kehitys Design Science Research-näkökulmasta. Kuva esittää, miten kohdeorganisaation liiketoimintatarpeet ohjaavat artefaktin kehitystä, joka perustuu tieteelliseen tietopohjaan ja tuottaa ratkaisuja master datan laadun parantamiseen. Kaavio pohjautuu Alan Hevnerin Design Science Research sykliin (Hevner, A., 2007).

Kuvio havainnollistaa tutkimuksen etenemistä iteratiivisena prosessina. Toimintaympäristöstä lähtevä nuoli yhdistää tutkimuksen kohdeorganisaation käytännön ongelmiin, tietoperustasta lähtevä nuoli aiempaan tutkimustietoon ja keskellä oleva sykli varsinaisen artefaktin kehittämiseen ja arviointiin. Malli korostaa sitä, että tekoälyratkaisua kehitettiin samanaikaisesti sekä teknisestä että organisatorisesta näkökulmasta.

3.2 Aineistonkeruumenetelmät

Tämän tutkimuksen aineisto kerättiin useista toisiaan täydentävistä lähteistä, jotta kohdeyrityksen master data -prosessien nykytila, käyttäjien tarpeet ja kehitettävän tekoälyratkaisun toimivuus voitiin arvioida monipuolisesti. Aineistonkeruu tukee myös tutkimuksessa toteutettua nykytila-analyysiä, jonka tavoitteena oli tunnistaa materiaalitiedon hallinnan keskeiset haasteet, tietovirtojen rakenteet sekä manuaaliset työvaiheet ennen tekoälyratkaisun suunnittelua. Menetelmien kokonaisuus noudattaa design science -tutkimuksen periaatteita, joissa yhdistyvät käyttäjälähtöinen ymmärrys, dokumentoitu evidenssi ja artefaktin kokeellinen validointi (Hevner ja muut, 2004).

Ensimmäinen aineistonkeruumenetelmä oli kohdeyrityksen master data -asiantuntijoille suunnattu kyselytutkimus, jonka tarkoituksena oli kartoittaa käyttäjien kokemuksia, prosessikohtaisia haasteita ja näkemyksiä uuden SAP-rikastustyökalun tarpeellisuudesta. Kysely sisälsi sekä strukturoituja Likert-asteikkoisia väittämiä että avoimia kysymyksiä, ja se toteutettiin sähköisenä lomakkeena. Kyselyyn osallistuneet edustivat yrityksen keskeisiä käyttäjäryhmiä, kuten suunnittelua, konfigurointia ja master datan ylläpitoa. Tämä mahdollisti eri käyttäjäroolien näkökulmien vertailun ja tarjosi laajemman käsityksen käyttäjävaatimuksista.

Toiseksi aineistona hyödynnettiin asiantuntijakeskusteluja ja projektipalavereita, joita käytiin kehitysvaiheiden aikana yrityksen SAP-asiantuntijoiden, suunnittelijoiden ja tuotetiedon ylläpitäjien kanssa. Näiden keskustelujen avulla täydennettiin käsitystä nimikekirjaston rakenteesta, datavirroista, rikastusprosessin kipupisteistä sekä tekoälyratkaisulle asetettavista toiminnallisista ja tietoturvaan liittyvistä vaatimuksista. Keskustelut tukivat myös artefaktin iteratiivista kehittämistä tarjoamalla reaaliaikaista palautetta ja tarkennuksia suunnitteluratkaisuihin.

Kolmantena aineistolähteenä hyödynnettiin kohdeyrityksen dokumentaatioita ja teknisiä tietoja. Näihin sisältyivät muun muassa materiaalitiedon ylläpidon ohjeistukset, SAP-järjestelmän Excel-pohjat, EPLAN Data Portalista haettu komponenttidata (kuten

kategoriat, tuoteryhmät ja tekniset attribuutit) sekä projektin sisäiset luonnokset arkkitehtuurista, dataputkista ja RAG-mallin rakenteesta. Dokumenttiaineisto tarjosi perustan nykytila-analyysille ja toimi teknisen ratkaisun suunnittelun lähtökohtana.

Neljänneksi tutkimuksessa kerättiin aineistoa tekoälyprototyypin tuottamista rikastus- ja hakutuloksista. Chatbotia testattiin rajatulla käyttäjäjoukolla, ja tekoälyn tuottamia Excel-muotoisia rikastusehdotuksia verrattiin yrityksen standardeihin ja olemassa oleviin SAP-tietoihin. Testaus tuotti olennaista tietoa artefaktin laadusta, luotettavuudesta ja soveltuvuudesta kohdeyrityksen prosesseihin.

Näiden menetelmien yhdistelmä tarjosi tutkimukselle kattavan aineistopohjan. Sen avulla voitiin arvioida sekä nykytila-analyysin havaintoja että kehitettävän tekoälyratkaisun soveltuvuutta kohdeyrityksen master data -ympäristöön. Tarkastelu kohdistui erityisesti tekniseen toimivuuteen ja käytännön hyötyyn.

3.3 Nykytila-analyysi

Nykytila-analyysi toteutettiin tutkimusmenetelmänä, jonka tarkoituksena oli muodostaa systemaattinen kokonaiskuva Arnon Oy:n master datan rakenteesta, laadusta ja hallintakäytännöistä. Menetelmän avulla arvioitiin, miten organisaation nykyinen datavaranto soveltuu tekoälypohjaisen eheytys- ja rikastusratkaisun kehittämiseen. Nykytila-analyysi auttoi myös keskeisten rakenteellisten ja sisällöllisten haasteiden tunnistamisessa, jotka tulisi joka tapauksessa korjata. Nykytila-analyysi suoritettiin alla olevan kaavion mukaan (Kuva 2).

Kuva 2. Prosessikaavio nykytila-analyysista. Analyysi koostui kuudesta vaiheesta, joista jokaisessa pyrittiin tuottamaan seuraavaa vaihetta avustava tuotos.

Ensimmäisessä vaiheessa kuvattiin nykyinen toimintamalli sekä tunnistettiin analyysin kannalta olennaiset tietolähteet. Tavoitteena oli muodostaa kokonaiskuva ilmiöstä ja siitä, mitä ongelmaa datalla ratkaistaan. Toisessa vaiheessa poistettiin analyysin kannalta epäolennaiset tiedot ja rajattiin aineisto tukemaan kehitettävää ratkaisua. Näin varmistettiin datan relevanssi ja laatu jatkokäsittelyä varten. Kolmannessa vaiheessa yhdistettiin eri lähteistä peräisin oleva tieto yhtenäiseksi kokonaisuudeksi yhteisten tunnisteiden avulla. Tämä mahdollisti tehokkaan käsittelyn ja vertailun. Neljännessä vaiheessa rajattiin aineisto tarkempaan tarkasteluun kehitysvaiheen tavoitteiden mukaisesti. Tässä keskityttiin tärkeimpiin tietoihin ja käyttötapauksiin. Viidennessä vaiheessa tunnistetut havainnot käytiin läpi yhdessä tiimin kanssa, ja niiden merkitystä sekä luotettavuutta arvioitiin käytännön näkökulmasta. Havaintojen pohjalta muodostettiin johtopäätökset. Viimeisessä vaiheessa suunniteltiin ratkaisun rakenne analyysin tulosten ja työprosessin ymmärryksen perusteella. Tässä huomioitiin myös tulevan työkalun keskeiset vaatimukset.

Nykytila-analyysi perustui laajaan materiaalidatan keräämiseen ja analysointiin useista toisiaan täydentävistä tietolähteistä. Näitä olivat esimerkiksi perustiedot, toimipistekohtaiset materiaalirivit, Classification-tuotetiedot, käyttödata sekä toimittajien tuotekuvaukset ja katalogit. Aineiston käsittelyssä hyödynnettiin iteratiivista analyysiprosessia, jossa lähdetaulujen tiedot yhdistettiin materiaalikoodien perusteella yhtenäiseksi analyysitaulukoksi. Tämä mahdollisti datan rakenteiden, sisältöjen ja niiden välisten riippuvuuksien tarkastelun kokonaisuutena.

Osana menetelmää suoritettiin datan esikäsittely ja siivous, jossa poistettiin analyysin tavoitteiden kannalta epäolennaiset tietueet. Tällä varmistettiin, että analyysi kohdistuu tuotannollisesti relevanttiin aineistoon ja että myöhemmin kehitettävät AI-mallit perustuvat laadukkaaseen sekä käyttöympäristöä vastaavaan dataan.

Nykytila-analyysin rakenne koostui datan laadun kuuden ulottuvuuden tarkastelusta: tarkkuus, ajantasaisuus, täydellisyys, yhdenmukaisuus, saatavuus ja uniikkius (Kuva 3).

Kuva 3. Datan laadun kuusi ulottuvuutta. Datan jakaminen eri ulottuvuuksiin antaa jäsennellyn tavan arvioida, onko data käyttökelpoista. Ulottuvuuksien avulla saadaan selkeämpi kuva tietoresurssien vahvuuksista ja heikkouksista (SAP, 2025).

Näiden avulla arvioitiin semanttisia tekijöitä, jotka vaikuttavat datan hyödynnettävyyteen tekoälypohjaisessa rikastamis- ja eheytysprosessissa. Analyysin tuottama kokonaiskuva muodosti perustan myöhemmille ratkaisumäärittelyille ja AI-mallien kehitystyölle.

3.4 Kyselytutkimus

Kyselytutkimus toteutettiin täydentävänä tutkimusmenetelmänä nykytila-analyysille, ja sen tarkoituksena oli kartoittaa master datan käyttäjien kokemuksia, tarpeita ja odotuksia tekoälyavusteista SAP-rikastamistyökalua kohtaan. Kyselyn avulla pyrittiin ymmärtämään sekä käytännön työn haasteita ostomateriaalien ja nimikkeiden parissa että niitä tekijöitä, jotka vaikuttavat uuden työkalun hyväksyttävyyteen ja käyttöönoton onnistumiseen.

Kohderyhmänä olivat Arnon Oy:n nimikekirjastoa ja ostomateriaaleja työssään hyödyntävät asiantuntijat. Vastaajat edustivat ensisijaisesti myyntiä ja asiakasrajapintaa sekä suunnittelua, minkä lisäksi mukana oli yksi vastaaja, jonka rooli sijoittui usean osa-alueen välimaastoon. Yhteensä vastauksia saatiin 13 kappaletta, mikä kattaa merkittävän osan nimikekirjaston aktiivisista käyttäjistä. Valtaosa vastaajista työskentelee ostomateriaalien ja nimikkeiden parissa päivittäin ja on käyttänyt nykyistä Falcony/Incy-työkalua nimikkeiden avaukseen yli kaksi vuotta. Näin varmistettiin, että vastaajilla on syvällinen ja käytännönläheinen ymmärrys nykyisistä prosesseista ja niiden haasteista.

Kysely toteutettiin verkkolomakkeena loka–marraskuussa 2025, ja vastaaminen oli vapaaehtoista. Vastaukset kerättiin anonyymisti siten, että yksittäisiä henkilöitä ei voida tunnistaa, mutta taustatekijöitä (esimerkiksi roolia ja käyttökokemusta) voidaan hyödyntää tulosten tulkinnassa. Kysely koostui sekä strukturoitujen että avoimien kysymysten yhdistelmästä. Strukturoituja kysymyksiä esitettiin muun muassa seuraavista teemoista:

- taustatiedot (rooli organisaatiossa, työskentelytiheys ostomateriaalien parissa, Falcony/Incy-kokemus)

- nimikekirjaston käyttötarkoitukset ja nykyiset haasteet
- olennaisiksi koetut materiaalitiedot ja nimikerakenteen selkeys
- odotukset tekoälyavusteisen rikastamistyökalun hyödyllisyydestä
- toiveet tulosten esitystavasta, käyttöliittymästä sekä koulutuksesta ja tuesta
- tekijät, jotka lisäävät luottamusta tekoälyn tuottamaan rikastettuun dataan.

Useissa kysymyksissä hyödynnettiin viisiportaista Likert-asteikkoa, jolla vastaajat arvioivat erilaisten ominaisuuksien ja toimintojen hyödyllisyyttä. Lisäksi useat kysymykset mahdollistivat usean vastausvaihtoehdon valitsemisen, jolloin pystyttiin tunnistamaan keskeisiä teema-alueita esimerkiksi haasteiden ja koulutustarpeiden osalta.

Kyselyn loppuun sisältyi avoimia kysymyksiä, joissa vastaajat saivat kuvata omin sanoin AI-työkalun käyttöönottoon liittyviä huolia ja riskejä. Lisäksi vastaajia pyydettiin arvioimaan, miten tekoälyavusteinen rikastus voisi tukea heidän työskentelyään sekä millaisia toiminnallisuuksia järjestelmään tulisi sisällyttää. Näiden tarkoituksena oli tuoda esiin sellaisia näkökulmia ja ideoita, jotka eivät välttämättä nouse esiin strukturoitujen kysymysten kautta.

Aineisto analysoitiin kahdella tasolla. Strukturoitujen kysymysten osalta tarkasteltiin vastausjakaumia ja keskityttiin erityisesti tyypillisimpiin valintoihin ja niiden suhteellisiin osuuksiin. Avoimet vastaukset analysoitiin laadullisen teemoihin jakamisen avulla eli vastaukset luokiteltiin toistuviin teemoihin, kuten datan laatuun, työprosessien sujuvuuteen, osaamisen säilymiseen ja luottamukseen liittyviin kysymyksiin. Näin kyselytutkimus tarjosi sekä kvantitatiivista että kvalitatiivista ymmärrystä käyttäjien tarpeista, asenteista ja huolenaiheista ja muodosti tärkeän lähtökohdan SAP-rikastamistyökalun suunnittelu- ja käyttöönottoprosessin määrittelylle.

4 Tulokset ja analyysi

Tässä luvussa esitetään ja analysoidaan tutkimuksen keskeiset tulokset, jotka on johdettu nykytilan analyysistä sekä henkilöstölle toteutetusta kyselytutkimuksesta. Tulokset tarjoavat kokonaisvaltaisen kuvan kohdeyrityksen master datan laadusta, siihen liittyvistä haasteista sekä käyttäjien kokemuksista ja odotuksista datan kehittämisen ja tekoälyratkaisujen näkökulmasta.

Luku on jäsennetty siten, että ensin syvennytään projektin tekniseen toteutukseen ja käydään projektissa käytetyt teknologiat läpi. Seuraavaksi tarkastellaan master datan nykytilaa objektiivisten laatuulottuvuuksien kautta. Nykytilan analyysissä keskitytään erityisesti datan tarkkuuteen, ajantasaisuuteen, täydellisyyteen, saatavuuteen, yhdenmukaisuuteen ja uniikkiuteen, jotka muodostavat keskeisen perustan datan hyödynnettävyydelle sekä automaation ja tekoälyn soveltamiselle. Näiden osa-alueiden analyysi pohjautuu järjestelmistä kerättyyn data-aineistoon ja sitä täydentävään tekniseen tarkasteluun.

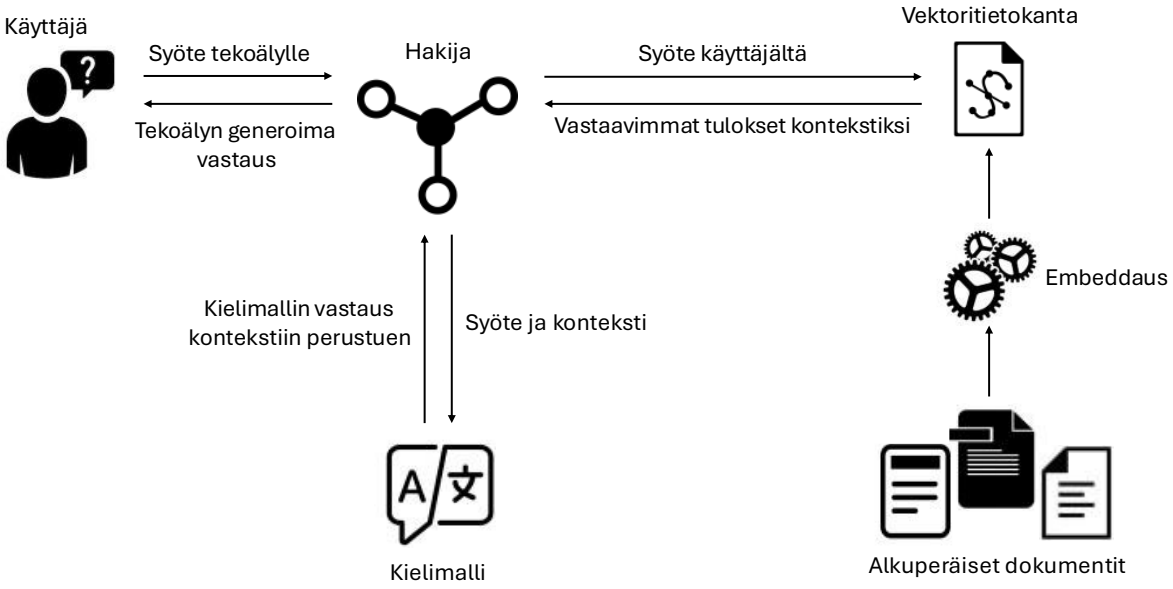
Tämän jälkeen luvussa käsitellään kyselytutkimuksen tulokset, joiden avulla tuodaan esiin käyttäjien näkemykset nimikekirjaston toimivuudesta, datan laadun vaikutuksista päivittäiseen työhön sekä odotuksista tekoälyavusteisia ratkaisuja kohtaan. Kyselytulokset täydentävät teknistä analyysiä tarjoamalla prosessi- ja käyttäjälähtöisen näkökulman, joka on keskeinen datan laadun kehittämisen onnistumiselle organisaatiotasolla.

Luvun kokonaisanalyysi luo perustan seuraavassa luvussa esitettävälle tekoälyavusteisen datan rikastamisen ja eheytyksen toteutuksen arvioinnille sekä johtopäätöksille ja kehitysehdotuksille.

4.1 Projektin tekninen toteutus ja artefaktin kuvaus

Projektin tekninen toteutus rakentuu Azure AI Foundryn ympärille ja noudattaa Design Science Research -viitekehyksen mukaista iteratiivista kehittämislogiikkaa. Tässä työssä DSR-menetelmän kolme keskeistä sykliä näkyvät selkeästi tutkimuksen eri vaiheissa. Kohdeyrityksen master dataan liittyvät käytännön ongelmat muodostavat relevanssisyklin lähtökohdan. Aiempi tutkimus ja tekoälyteknologian vakiintunut tietopohja puolestaan muodostavat rigor-syklin perustan (Hevner ja muut, 2004). Artefaktin suunnittelu, toteutus ja arviointi muodostavat design-syklin keskeisen sisällön. Tämä kolmijako tukee sitä, että ratkaisua kehitetään sekä käytännön tarpeen että tieteellisen perustan näkökulmasta.

Ratkaisussa hyödynnetään Retrieval-Augmented Generation (RAG) -arkkitehtuuria, jossa kielimallin generatiivinen kyvykkyys yhdistetään ulkoisesta tietovarannosta haettuun kontekstiin. Microsoft Foundryn dokumentaation mukaan RAG on malli, jossa haku ja generointi toimivat yhdessä niin, että vastaukset voidaan ankkuroida organisaation omaan dataan, indeksiin ja grounding-dataan. Tämän lähestymistavan etuna on se, että yrityksen sisäistä tietoa voidaan hyödyntää ilman, että koko tietopohja täytyisi upottaa kielimallin parametreihin raskaan hienosäädön avulla (Microsoft, 2026). RAG-mallin vastaukset sisältävät olennaista tietoa ja ne voivat sisältää viittauksia lähteisiin (Microsoft, 2026). RAG-mallin toiminta käyttäjän näkökulmasta esitetty kuvassa 4 (Kuva 4).



Kuva 4. RAG-arkkitehtuuri ja tiedonkulku. Käyttäjän tehtyä kyselyn, hakijamalli hakee indeksoidusta tietokannasta kysymystä vastaavimman sisällön ja lähettää sen kielimallille kontekstiin yhdistettynä. Lopuksi kielimalli luo vastauksen haetun sisällön ja käyttäjän kyselyn perusteella.

RAG-malli perustuu kolmivaiheiseen prosessiin:

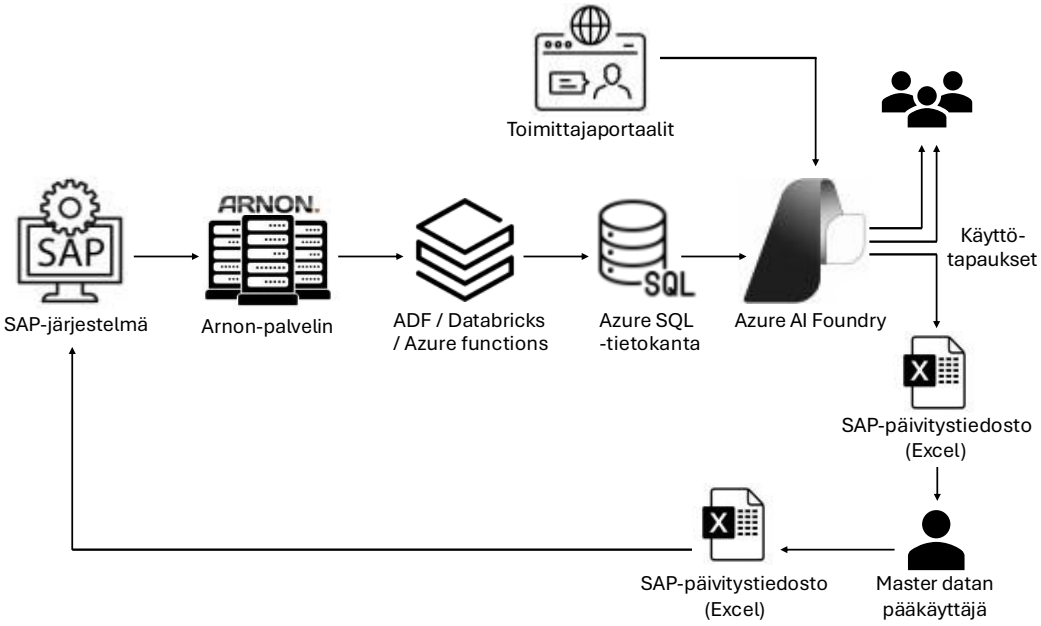
1. Haku (Retrieve): Kun käyttäjä esittää kysymyksen, hakija tekee kyselyn vektoritietokantaan, josta se etsii vastaavimmat ja asiaankuuluvimmat tulokset.
2. Syötteen täydennys (Augment): Sovellus yhdistää käyttäjän kysymyksen ja noudetut tulokset (grounding-data) kehotteeksi, jossa tiedot on yhdistetty kontekstiin.
3. Vastauksen luonti (Generate): Kielimalli vastaanottaa täydennetyn kehotteen ja muodostaa vastauksen, joka perustuu käyttäjän kysymykseen ja noudettuihin tuloksiin. Vastaus luodaan siis vain näiden pohjalta, mikä vähentää epätarkkuuksia ja mahdollistaa tarkat viittaukset lähteisiin.

RAG-malli toimii parhaiten, kun tiedonhaku onnistuu nopeasti ja johdonmukaisesti. Tämän takia alkuperäiset dokumentit kannattaa muuntaa sellaiseen muotoon, että hakija saa etsittyä asiaankuuluvan sisällön mahdollisimman nopeasti ja pienillä

resursseilla (Microsoft 2026). Alkuperäiset dokumentit siis indeksoidaan ja muunnetaan vektorimuotoon, jotta malli toimii tehokkaasti

4.1.1 Tekninen arkkitehtuuri ja tiedonkulku

Suunniteltu tekninen tiedonkulku perustuu pilvipohjaiseen data- ja tekoälyarkkitehtuuriin, jossa master dataa käsitellään useiden integraatioon ja analytiikkaan liittyvien komponenttien kautta. Tämän arkkitehtuurin tarkoituksena on mahdollistaa datan käsittely, laadunhallinta sekä tekoälyohjainen analysointi ja rikastaminen. Järjestelmän tiedonkulku ja keskeiset komponentit on esitetty kuvassa 5 (Kuva 5).



Kuva 5. Järjestelmän tekninen arkkitehtuuri ja tiedonkulku järjestelmien välillä. Arkkitehtuuri on suunniteltu mahdollistaen datan käsittelyn, laadunhallinnan sekä tekoälypohjaisen analysoinnin ja rikastamisen. Viimeisen validoinnin suorittaa master datan pääkäyttäjä.

Tiedonkulku toimii siten, että data siirtyy SAP-järjestelmästä kohdeyrityksen integroidun palvelimen kautta Azure-ympäristöön. Seuraavaksi dataa käsitellään Azure Data Factoryn (ADF) ja Databricksin avulla. Nämä muuntavat datan optimoituun ja koneluettavaan

muotoon, jotta RAG-malli toimii tehokkaammin. Käsitelty data tallennetaan Azure SQL -tietokantaan, joka toimii tekoälytyökalun tietovarastona.

Data siirtyy Azure SQL -tietokannasta Azure AI Foundryyn, johon tekoälytyökalu on rakennettu. Azure AI Foundry mahdollistaa sen, että työkalulla pystytään analysoimaan, validoimaan, eheyttämään sekä rikastamaan master dataa. Toimittajaportaaleista tulevat tiedot ovat ulkoisia lähteitä, joita työkalu voi käyttää tiedon rikastamiseen.

Käyttäjien suoritettua käyttötapauksia, työkalu muodostaa SAP-päivitystiedoston (Excel) rikastettavien tietojen pohjalta, jonka master datan pääkäyttäjä validoi ja lisää tietokantaan.

4.1.2 Arkkitehtuuri ja hybridihaku teollisessa mittakaavassa

Järjestelmän tekninen suorituskyky on mitoitettu vastaamaan kohdeyrityksen laajan nimikekirjaston vaatimuksia. Nykytila-analyysissä tarkasteltu aineisto käsittää noin 60 000–70 000 nimikettä, mikä asettaa korkeat vaatimukset haun tarkkuudelle ja relevantin tiedon löytämiselle. Tästä syystä ratkaisussa hyödynnetään hybridihakua, joka yhdistää avainsanapohjaisen haun ja vektorihaun edut. Azure AI Searchin dokumentaation mukaan hybridihaku perustuu tekstihakujen ja vektorihakujen rinnakkaiseen suorittamiseen (Microsoft, 2026). Hakutulokset yhdistetään Reciprocal Rank Fusion (RRF) -menetelmällä, jolloin järjestelmä hyödyntää sekä tarkkoja hakuosumia että semanttista samankaltaisuutta.

RRF-menetelmä yhdistää samanaikaisesti monta hakumenetelmää ja muodostaa niiden tuloksien perusteella yhden paremman hakutuloslistan. Menetelmän toiminta perustuu siihen, että dokumentit, jotka esiintyvät tai sijoittuvat korkealle useissa hauissa, ovat todennäköisesti relevanteimpia.
RRF-menetelmä toimii näin:

1. Useita hakuja suoritetaan rinnakkain.
2. Jokainen dokumentti pisteytetään sen sijoituksen perusteella.

3. Dokumentin pisteet lasketaan yhteen eri hakumenetelmistä.
4. Dokumentit järjestetään yhteenlasketun pistemäärän avulla lopulliseen järjestykseen.

Lopputuotos on yhdistetty hakutuloslista, jossa dokumentit on listattu järjestykseen pistemäärän perusteella. Paljon pisteitä saaneet dokumentit sijoittuvat korkeammalle. (Microsoft, 2026)

Järjestelmän dataputket perustuvat Azure AI Foundryn ja Azure AI Searchin ympäristöön rakennettuihin hakurakenteisiin ja indeksointimekanismeihin. Samassa indeksissä voidaan hyödyntää sekä tekstimuotoista sisältöä että vektorimuotoisia semanttisia esityksiä. Tällainen rakenne mahdollistaa sekä perinteisten hakuehtojen että semanttisen samankaltaisuuden hyödyntämisen haussa. Tämä on erityisen tärkeää teknisessä nimikekirjastossa, jossa tuotteiden nimet, koodit ja ominaisuudet voivat olla hyvin samankaltaisia, vaikka ne eivät ole täysin identtisiä. (Microsoft, 2026)

Kielimallin toimintaa ohjataan erillisellä ohjetiedostolla, joka sisältää tarkat säännöt sille, miten hakua suoritetaan ja miten löydettyä tietoa tulee painottaa suhteessa käyttäjän kyselyyn. Foundryn dokumentaatio korostaa, että RAG-järjestelmissä promptit, järjestelmäviestit ja grounding-data ohjaavat mallia käyttämään haettua tietoa perustana vastaukselle. Tämän vuoksi ohjeistuksen laatu on keskeinen osa artefaktin toimivuutta, sillä se vaikuttaa suoraan siihen, kuinka hyvin malli hyödyntää haettua kontekstia ja kuinka johdonmukaisia vastaukset ovat.

4.1.3 Monilähteinen datan haku ja rikastaminen

Tekoälyratkaisun tietopohja on rakennettu useista lähteistä, jotka yhdessä muodostavat rikastusehdotusten kontekstin. Keskeinen periaate on, että RAG-järjestelmä hakee relevantin grounding-datan indeksistä ja käyttää sitä mallin vastauksen pohjana. Tässä työssä tämä periaate konkretisoituu yrityksen sisäisten komponenttitietokantojen, EPLAN Data Portalin sekä rajatun verkkohaun yhdistelmänä. Tavoitteena on tuottaa

rikastusehdotuksia, jotka ovat mahdollisimman hyvin linjassa sekä organisaation sisäisen master datan että ulkoisten teknisten lähteiden kanssa.

EPLAN Data Portalin osalta hyödynnetään kahta eri menetelmää. Komponenttien kategoriat on noudettu skriptipohjaisesti, mutta tekoäly kykenee myös suorittamaan dynaamisia hakuja suoraan portaalin tiedoista. Tämä yhdistetty tietopohja varmistaa, että tekoäly voi muodostaa tarkempia rikastus- ja eheytysehdotuksia kuin pelkän sisäisen aineiston perusteella olisi mahdollista. Monilähteinen haku tukee erityisesti tilanteita, joissa yksittäinen lähde ei sisällä kaikkia tarvittavia teknisiä attribuutteja.

4.1.4 Prompt engineering ja vastausten validointi lähteiden avulla

Tekoälyn tuottamien vastausten tarkkuutta ja luotettavuutta on parannettu määrittelemällä sille tarkat vastausmuodot ja käyttöehdot. RAG-järjestelmissä promptit, system message -ohjeistus ja grounding-data toimivat yhdessä siten, että malli tuottaa vastauksen haetun tiedon pohjalta eikä pelkästään mallin sisäisen muistin varassa. Tässä työssä tämä näkyy template-pohjaisessa vastausrakenteessa, jossa vastauksen muoto on vakioitu ja tieto esitetään käyttäjälle mahdollisimman järjestelmällisesti.

Läpinäkyvyyden lisäämiseksi työkalu merkitsee käyttämänsä lähteet vastausten yhteyteen. Käyttäjä näkee siten, onko tieto peräisin Arnonin omasta tietokannasta, EPLAN-lähteistä vai rajatusta verkkohauusta. Tämä on tärkeää, koska RAG-arkkitehtuurin keskeinen vahvuus on juuri se, että vastaukset voidaan ankkuroida lähdedataan ja siten vähentää perustelemattomien tai virheellisten vastausten riskiä. Käytännössä lähdeviitteet tukevat myös käyttäjän mahdollisuutta arvioida rikastusehdotuksen soveltuvuutta ennen lopullista hyväksyntää.

4.1.5 Web-sovellus ja hallittu SAP-integraatioprosessi

Tekoälytyökalu toteutettiin chatbot-ratkaisuna, joka toimii erillisenä web-sovelluksena. Alkuperäiseen suunnitelmaan sisältyi integraatio Microsoft Teams -ympäristöön, mutta

testausvaiheessa ilmenneiden teknisten haasteiden vuoksi web-sovellus todettiin toimintavarmemmaksi vaihtoehdoksi. Työkalu on suunniteltu tukemaan suunnittelijoita, myyjiä ja master data -asiantuntijoita heidän päivittäisessä työssään tarjoamalla nopean pääsyn materiaalitietoihin.

Tekoälyn tuottamat rikastukset toimitetaan Excel-muotoisina taulukoina. Tämä on valittu siksi, että taulukot ovat yhteensopivia SAP-järjestelmän olemassa olevan lataustyökalun kanssa. Prosessi noudattaa hallittua toimintamallia, jossa teknologia ei tee suoria muutoksia tietokantaan. Lopullinen vastuu tietojen oikeellisuudesta ja tallentamisesta SAP-järjestelmään on master datan pääkäyttäjällä, joka toimii prosessin tarkastajana ja hyväksyjänä.

4.1.6 Kehityksen nykytila ja iteraatio

Kokonaisuutta kehitetään Design Science Research -viitekehyksen mukaisesti iteratiivisesti. Tämä tarkoittaa, että dataputkia, hakumallia ja chatbotin toiminnallisuuksia parannetaan vaiheittain käyttäjäpalautteen ja teknisen suorituskyvyn perusteella. DSR-kirjallisuudessa korostetaan juuri tätä syklistä etenemistä. Artefaktia rakennetaan ja arvioidaan toistuvasti, sen toimivuutta verrataan ympäristön tarpeisiin, ja samalla ratkaisu ankkuroidaan tutkimuksen tietopohjaan. Vaikka varsinaista lopullista validointia ei ole vielä täysin suoritettu testivaiheen ollessa kesken, on ohjeiden ja vastausmallien jatkuva kehitys jo nyt parantanut työkalun tuottamien tulosten laatua.

4.1.7 Arkkitehtuurin rajaukset ja tiedonlähteiden hallinta

Tässä tutkimuksessa kehitetyn RAG-arkkitehtuurin keskeinen periaate on, että generoitavat vastaukset perustuvat ensisijaisesti organisaation hyväksyttyihin ja hallittuihin tietolähteisiin. Näihin kuuluvat yrityksen sisäinen master data, EPLAN Data Portal sekä muut erikseen määritellyt tekniset tietolähteet. Vaikka kehitysvaiheessa on

kokeiltu myös rajattua verkkohakua (esimerkiksi Bing Search -rajapinnan kautta), tätä ei ole tarkoitettu osaksi lopullista tuotantoratkaisua.

Verkkohaku toimii tässä työssä eksploratiivisena tukimekanismina, jonka avulla voidaan arvioida ulkoisen tiedon potentiaalia rikastamisessa tilanteissa, joissa sisäinen data on puutteellista. Tällainen lähestymistapa on linjassa RAG-kirjallisuuden kanssa, jossa korostetaan kontrolloidun tietopohjan merkitystä faktuaalisen tarkkuuden ja luotettavuuden varmistamisessa (Lewis ja muut, 2020).

Lopullisessa tuotantoympäristössä tiedonlähteet tulee rajata organisaation hyväksymiin ja validoituihin lähteisiin. Tavoitteena on vähentää virheellisen tai epäluotettavan tiedon riskiä sekä varmistaa datan hallittavuus, tietoturva ja jäljitettävyys.

4.2 Nykytilan analyysin tulokset

Nykytila-analyysin tavoitteena oli muodostaa kokonaiskuva Arnon Oy:n master datan rakenteesta, laadusta ja hallintaprosesseista sekä arvioida sen soveltuvuutta tekoälypohjaisen datan rikastamis- ja eheytysratkaisun kehittämiseen. Analyysi toteutettiin osana laajempaa Master Data AI -hanketta, jossa keskityttiin materiaalihallinnan ja tuoterakenteiden (Bill of Materials, BOM) tietojen automatisoituun käsittelyyn.

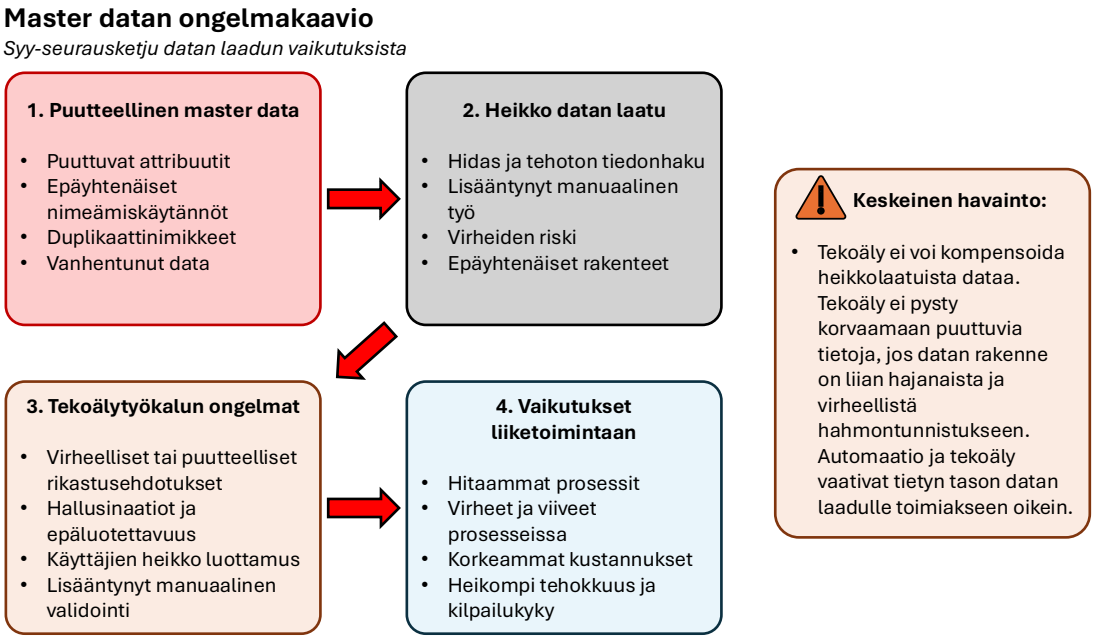
Nykytila-analyysin lähtökohtana oli käsitellä Arnonin materiaalidataa sellaisessa laajuudessa, että sen perusteella voitiin tehdä luotettava arvio datan laadullisista ominaisuuksista ja rakenteellisista haasteista. Käytettävissä olleet pääasialliset tietolähteet olivat:

- Materiaalidata, joka sisälsi sekä perustiedot että toimipistekohtaiset rivit
- Classification-tuotetiedot
- Käyttödata Tampereen toimipisteestä
- Toimittajakohtaiset tuotekuvaukset ja katalogit, joita hyödynnettiin tiedon rikastamisen arvioinnissa

Datan käsittelyssä käytettiin iteratiivista menetelmää, jossa eri lähdetaulujen tiedot yhdistettiin materiaalikoodin avulla yhtenäiseksi analyysitaulukoksi. Tämä mahdollisti datan kokonaisvaltaisen tarkastelun sekä sen rakenteiden ja sisältöjen välisten yhteyksien arvioinnin. Analyysi rajattiin niihin toimittajiin, joiden osuus käytetyistä materiaaleista oli merkittävin.

Datan esikäsittely sisälsi myös materiaalin siivouksen, jossa poistettiin kehitystyön kannalta epäolennaiset tiedot, kuten käytöstä poistettujen toimipaikkojen materiaalit sekä Arnonin itse valmistamat tuotteet. Tämä vaihe oli kriittinen, jotta analyysiin jäävä data vastaisi realistisesti tuotannossa käytettävää aineistoa ja jotta tekoälymallin koulutuksessa ei hyödynnettäisi virheellisiä tai epäolennaisia tietueita.

Nykytila-analyysin yhteydessä arvioitiin datan laadun kuusi keskeistä ulottuvuutta: tarkkuus, ajantasaisuus, täydellisyys, yhdenmukaisuus, saatavuus ja uniikkius. Näiden ulottuvuuksien perusteella muodostettiin kokonaiskuva siitä, missä määrin olemassa oleva data täyttää tekoälyratkaisujen kehittämiselle asetetut tekniset ja semanttiset vaatimukset. Master datan ongelmakaavio havainnollistaa datan laadun vaikutusta yrityksen liiketoimintaan (Kuva 6).



Kuva 6. Master datan ongelmakaavio. Kaavio havainnollistaa master datan huonon laadun vaikutuksen tekoälytyökalujen toimintaa ja liiketoimintaan. Keskeisenä havaintona huomataan, että tekoälytyökalut ja automaatio vaativat tietyntasoista dataa, jotta ne toimivat halutulla tavalla.

Kaavio havainnollistaa, miten master datan puutteet vaikuttavat vaiheittain tekoälytyökalun suorituskykyyn ja yrityksen liiketoimintaprosesseihin. Puuttuvat attribuutit, epäyhtenäiset nimeämiskäytännöt ja duplikaattitiedot heikentävät datan laatua, mikä lisää manuaalista työtä ja virheiden riskiä. Tämän puolestaan heikentää tekoälyratkaisun luotettavuutta ja vähentää automaation hyötyjä. Heikkolaatuinen master data aiheuttaa siis ketjureaktion, joka heikentää koko AI-ratkaisun toimintaa.

4.2.1 Tarkkuus ja ajantasaisuus

Materiaalidatan tarkkuuden arviointi osoittautui haastavaksi datan suuren volyymin ja monimutkaisen rakenteen vuoksi. Tiedot kuvaavat laajaa materiaalivalikoimaa, eikä kaikkien tietueiden semanttista oikeellisuutta ollut mahdollista vahvistaa ilman kontekstuaalista asiantuntemusta. Ajantasaisuuden osalta havaittiin, että SAP-järjestelmässä tapahtuvat päivitykset ovat pitkälti manuaalisia ja riippuvaisia erillisistä päivitysprosesseista. Tietokantoihin data päivittyy automaattisesti yöajoina, mutta

muutokset eivät välity reaaliajassa muihin järjestelmiin, mikä heikentää datan operatiivista käyttökelpoisuutta.

4.2.2 Täydellisyys ja saatavuus

Analyysi paljasti puutteita erityisesti MS Book Part No -kentässä. Useilta materiaaleilta puuttui kyseinen tieto, vaikka se olisi pääteltävissä Material Description -kentän perusteella. Lisäksi havaittiin rivejä, joissa sekä tuotteen kuvaus että osanumero olivat tyhjiä. Data on tällä hetkellä tallennettuna Arnonin omalle palvelimelle, mikä rajoittaa sen saavutettavuutta ja vaikeuttaa integraatiota nykyaikaisiin pilvipohjaisiin dataratkaisuihin. Pilvimigraatio Azure SQL -tietokantaan on kuitenkin suunniteltu seuraavaan kehitysvaiheeseen, mikä parantaa tiedon saatavuutta ja prosessointiautomatiikkaa.

4.2.3 Yhdenmukaisuus ja uniikkius

Merkittävä osa havaituista ongelmista liittyi tietojen epäyhtenäisyyteen. Saman toimittajan nimeä esiintyi useissa eri muodoissa, mikä vaikeuttaa tiedon yhdistämistä ja tekoälypohjaista päättelyä. Lisäksi poistettujen materiaalien merkintätavat olivat epästandardisoituja. Käytössä oli useita eri muotoja kuten “DO NOT USE”, “DO NOT USE xxxx” ja “DON’T USE”. Tämä aiheuttaa epävarmuutta datan tulkinnassa ja automatisoidussa luokittelussa.

Uniikkiusanalyysi paljasti useita potentiaalisia duplikaattitapauksia, joissa identtiset Material Description- ja MS Book Part No -tiedot esiintyivät eri Material ID -tunnuksilla tai hinnoilla. Osa tapauksista johtuu kuitenkin tuotteiden tarkoituksellisista variaatioista, kuten eri pakkauskoosta tai toimitusmuodosta, joten yksiselitteinen automaattinen duplikaattien poisto ei ole teknisesti perusteltua ilman manuaalista validointia.

4.2.4 Data-analyysin yhteenveto

Yhteenvetona voidaan todeta, että nykytila-analyysi osoitti kohdeyrityksen master datan olevan rakenteellisesti monipuolinen mutta laadullisesti epätasainen. Datan eheys ja standardointi vaativat systemaattisia toimenpiteitä, jotta se soveltuu luotettavasti tekoälypohjaisen rikastamisen ja automaattisen päätöksenteon perustaksi. Keskeisiksi kehityskohteiksi tunnistettiin toimittajatietojen harmonisointi, materiaalikenttien standardointi sekä tiedonhallintaprosessien automatisointi. Näiden toimenpiteiden kautta voidaan parantaa datan semanttista yhtenäisyyttä ja mahdollistaa tehokas tekoälyratkaisujen käyttöönotto osana kohdeyrityksen dataekosysteemiä.

4.3 Kyselytutkimuksen tulokset

Tässä luvussa esitetään master datan käyttäjille toteutetun kyselytutkimuksen keskeiset tulokset sekä niiden analyysi. Kyselyn tavoitteena oli selvittää käyttäjien kokemuksia nykyisistä master datan käsittelyprosesseista sekä tunnistaa työssä esiintyviä haasteita. Lisäksi tavoitteena oli kartoittaa käyttäjien odotuksia ja mahdollisia huolia uuden tekoälypohjaisen SAP-rikastamistyökalun käyttöönottoon liittyen. Kyselyssä hyödynnettiin henkilöstölle tuttuja termejä ja käsitteitä, jotta vastaaminen olisi mahdollisimman sujuvaa ja vastaajat tunnistaisivat omat työtilanteensa kysymyksistä.

Tulokset on analysoitu sekä määrällisesti että laadullisesti. Kvantitatiivinen analyysi tarjoaa kokonaiskuvan vastausten jakautumisesta, kun taas avoimet vastaukset tuottavat syvällisempää ymmärrystä käyttäjien kokemuksista, toiveista ja riskeihin liittyvistä näkökulmista. Näiden yhdistelmä muodostaa selkeän pohjan työkalun suunnittelulle, kehitykselle ja käyttöönottostrategialle.

4.3.1 Vastaajien taustat ja nimikekirjaston käyttö

Kyselyyn vastasi 13 henkilöä, joista kahdeksan (61,5 %) työskenteli myynnissä tai asiakasrajapinnassa ja neljä (30,8 %) suunnittelussa. Yksi vastaaja kuvasi rooliaan

useamman osa-alueen yhdistelmänä. Tämä jakauma heijastaa hyvin niitä pääasiallisia käyttäjäryhmiä, joiden työn kannalta master data ja nimikekirjasto ovat keskeisiä.

Suurin osa vastaajista työskentelee ostomateriaalien ja nimikkeiden parissa päivittäin (69,2 %), ja lisäksi 23,1 % tekee tätä työtä viikoittain. Kyselyn perusteella nimikekirjastoa hyödynnetään erityisen intensiivisesti myynti- ja suunnitteluprosesseissa, koska kaikki vastaajat ilmoittivat käyttävänsä nimikekirjastoa suunnitteluun ja tuoterakenteiden laatimiseen, ja valtaosa myös myyntiin ja tarjoustyöhön. Materiaalitietojen ylläpito, ostot/toimitusketju sekä varastonhallinta nousivat lisäksi esiin täydentävinä käyttökohteina.

Kokemus nykyisestä Falcony/Incy-työkalusta oli vastaajilla tyypillisesti pitkä. Yli 80 % vastaajista oli käyttänyt työkalua nimikkeiden avaukseen enemmän kuin kaksi vuotta, ja vain yksittäiset vastaajat olivat käyttäneet sitä alle kaksi vuotta. Tämä taustatekijä vahvistaa tulkintaa siitä, että vastaajat tuntevat nykyiset prosessit ja järjestelmät hyvin ja pystyvät arvioimaan niiden vahvuuksia ja heikkouksia perustellusti.

4.3.2 Keskeiset haasteet nimikekirjaston käytössä

Vastaajia pyydettiin tunnistamaan, mitä he pitävät haastavimpina tai työläimpinä tehtävinä nimikekirjaston parissa. Eniten mainintoja saivat korvaavuuksien etsiminen sekä nimikkeiden kuvaus ja nimeäminen. Näiden lisäksi haastaviksi koettiin hyväksyntä- ja standarditietojen varmistaminen, tuotekoodien selvittäminen, teknisten tietojen kerääminen sekä valmistajan ja MPN-tunnisteiden varmistaminen. Vastaukset havainnollistettu kuvassa 7 (Kuva 7).

Kuva 7. Kyselytutkimuksen vastaukset haasteista. Haastavimmaksi koettiin korvaavuuksien etsiminen ja kuvaus/nimeäminen. Yksi vastaaja lisäsi oman vastauksen ”Nimikekirjastoon lisättyjen tuotteiden nimitys ja lisätiedot ovat usein todella puutteellisia ja/tai harhaanjohtavia”.

Haasteet eivät rajoittuneet yksittäisiin teknisiin ongelmiin, vaan kuvastivat laajempaa rakenteellista kokonaisuutta:

- **Tiedon hajanaisuus ja löydettävyys:** Nimikkeisiin liittyvä tieto sijaitsee useissa eri lähteissä, ja tuotekoodien, MPN- ja valmistajatiedon varmistaminen edellyttää usein useiden sivustojen ja dokumenttien läpikäyntiä.
- **Manuaalinen ja aikaa vievä työ:** Materiaalien ja nimikkeiden avaaminen, tietojen tarkistaminen ja korvaavuuksien vertailu kuvattiin toistuvasti manuaaliseksi ja rutiininomaiseksi työksi, joka vie merkittävästi aikaa varsinaiselta asiantuntijatyöltä.
- **Teknisten tietojen ja dokumentaation puutteet:** Vastauksissa tuotiin esiin, että teknisiä speksejä ja dokumentaatiota (esim. datasheetit, hyväksynnät, standardit) on vaikea löytää tai ne ovat epäyhtenäisiä.
- **Nimeämisen ja kuvauskäytäntöjen epäyhtenäisyys:** Nimikkeiden nimeämisessä ja kuvaamisessa koettiin paljon vaihtelua. Ohjeita ei aina tunneta tai niitä sovelletaan eri tavoin, mikä hankaloittaa nimikkeiden vertailua ja hakua.

- **Työkalujen ja prosessien monimutkaisuus:** Falcony/Incy koettiin joissakin vastauksissa hitaaksi tai monimutkaiseksi, mikä lisää virheiden riskiä ja tekee nimiketietojen hallinnasta kognitiivisesti kuormittavaa.

Yhteenvetona voidaan todeta, että nykyiset haasteet liittyvät sekä datan laatuun ja yhtenäisyyteen että prosessien ja työkalujen käytettävyyteen. Nämä havainnot muodostavat perustelun tekoälyavusteisen rikastustyökalun kehittämiselle. Keskeisenä tavoitteena on minimoida manuaalista työtä, eheyttää hajanaista tietoa sekä yhtenäistää nimikkeiden käsittelyprosesseja.

4.3.3 Olennaiset materiaalitiedot ja rakenteelliset odotukset

Kyselyssä kartoitettiin myös, mitkä materiaalitiedot käyttäjät kokevat olennaisimmiksi. Eniten mainintoja saivat tuotekoodi tai tunniste, hinta ja toimitusaika, joita piti tärkeinä lähes kaikki vastaajat. Näiden lisäksi korostuivat valmistajan tiedot, tuotekuvaus sekä hyväksynnät ja standardit. Myös tekniset speksit (esimerkiksi jännite, virta, teho, lämpötila-alue, IP-luokka) nousivat toistuvasti esiin.

Tulokset osoittavat, että käyttäjille keskeisiä ovat sekä kaupalliset (hinta, toimitusaika) että tekniset ja laadulliset attribuutit (valmistaja, tekniset speksit, hyväksynnät). Master datan rikastamisen kannalta tämä tarkoittaa, että tekoälytyökalun tulee pystyä kokoamaan ja tarkistamaan sekä määrällistä että laadullista tietoa luotettavasti ja johdonmukaisesti.

Rakenteellisten odotusten osalta vastaajat olivat hieman kahtiajakautuneita. Hieman yli puolet vastaajista katsoi, että materiaalit ja nimikkeet tulisi luokitella kategorioihin ominaisuuksiensa mukaan, kun taas merkittävä vähemmistö ei pitänyt lisäkategorisointia tarpeellisena. Kyselytulokset nimikerakenteen selkeydestä osoittavat selvän kehitystarpeen. Valtaosa vastaajista piti nykyistä rakennetta ainoastaan kohtalaisena tai jopa epäselvänä, ja vain yksittäinen vastaaja arvioi sen olevan selkeä ja riittävä (Kuva 8).

Kuva 8. Kyselytutkimuksen vastaukset nimikkeiden nykyisen rakenteen selkeydestä. Valtaosa pitää nykyistä rakennetta epäselkeänä tai puutteellisena.

Tämä viittaa siihen, että nimikerakenteen harmonisointi ja läpinäkyvämpi hierarkia ovat keskeisiä kehitysalueita. Tekoälytyökalulta odotetaan tältä osin ennen kaikkea kykyä tuottaa yhdenmukaisia ja vertailukelpoisia nimike- ja attribuuttirakenteita, ei pelkästään lisätä yksittäisiä kenttiä.

4.3.4 Odotukset tekoälytyökalun hyödyllisyydestä

Tekoälypohjaisen rikastamistyökalun hyödyllisyyttä tarkasteltiin useista näkökulmista. Vastaajia pyydettiin arvioimaan viisiportaisella asteikolla (1 = ei lainkaan hyödyllinen, 5 = erittäin hyödyllinen) eri toimintojen merkitystä.

- **Automaattiset rikastusehdotukset tuotekoodin perusteella:** Valtaosa vastaajista arvioi tämän toiminnon erittäin tai melko hyödylliseksi: suurin osa antoi arvosanan 4, ja osa korkeimman arvosanan 5. Kukaan ei arvioinut toimintoa hyödyttömäksi.
- **Työkalun oppiminen käyttäjän tekemistä korjauksista:** Myös kyky mukautua käyttäjäkorjausten perusteella arvioitiin erittäin myönteisesti. Suurin osa

vastaajista piti tätä ominaisuutta vähintään melko hyödyllisenä, ja osa arvioi sen erittäin hyödylliseksi.

- **Rikastusehdotusten perusteleminen:** Ehdotusten läpinäkyvyys ja perustelut koettiin tärkeiksi, mutta arvioissa oli hieman enemmän hajontaa. Vaikka moni antoi tälle ominaisuudelle korkeita arvosanoja, osa vastaajista suhtautui siihen neutraalimmin. Tämä voi kertoa siitä, että kaikki eivät vielä hahmota konkreettisesti, millainen “perusteltu ehdotus” olisi käytännön työssä, vaikka periaatteellinen tarve läpinäkyvyydelle tunnistetaan.

Avoimet vastaukset tukevat kvantitatiivista kuvaa. Tekoälyavusteisen rikastamisen odotetaan erityisesti:

- vähentävän manuaalista tiedon hakua useista eri lähteistä
- nopeuttavan hinnoittelua ja komponenttivalintaa suunnittelussa
- helpottavan myytyjen tuotteiden rakenteiden käsittelyä ja hinnoittelua
- parantavan nimikkeiden yhtenäisyyttä ja vähentävän duplikaatteja
- tukevan asiakkaan palvelua, kun oikea komponentti ja sen korvaavuudet löytyvät nopeammin.

Kokonaisuutena vastaajat näkevät tekoälytyökalun ennen kaikkea työn tehostajana ja laadun parantajana, ei olemassa olevien asiantuntijaroolien korvaajana.

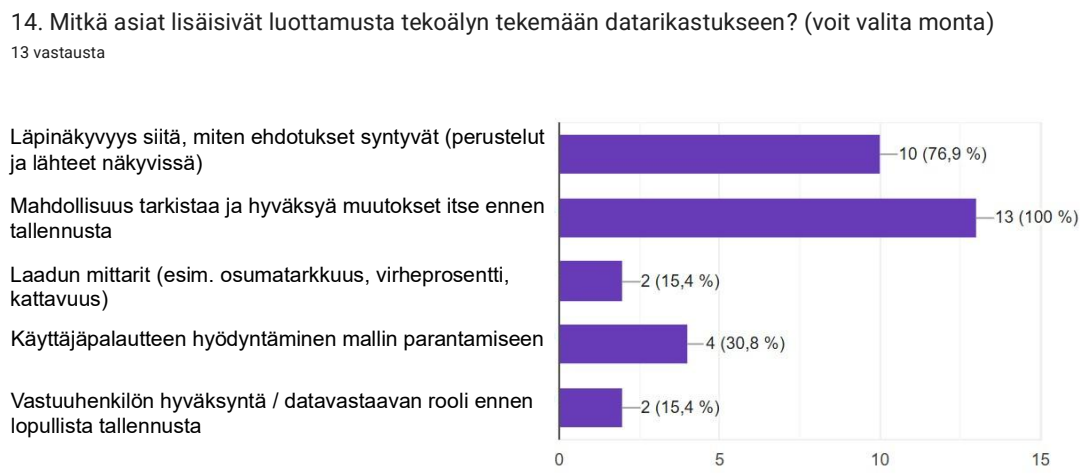
4.3.5 Tulosten esitystapa, käyttöliittymä ja käyttöönoton tuki

Tulosten esitystapaa koskevissa kysymyksissä nousi selkeä preferenssi vertailtavuuden ja muokattavuuden suuntaan. Selkeimpänä esitystapana pidettiin näkymää, jossa nykyinen tieto ja työkalun ehdottama rikastettu tieto esitetään rinnakkain. Monet vastaajat toivoivat lisäksi taulukkomuotoista, Excel-tyyppistä näkymää sekä interaktiivisia ominaisuuksia, joiden avulla käyttäjä voi suoraan korjata ja hyväksyä tietoja. Toisin sanoen toiveet kohdistuvat järjestelmään, jossa AI ei vain “kirjoita päälle”, vaan käyttäjä toimii aktiivisena päätöksentekijänä ja laadunvarmistajana.

Käyttöönoton tuen osalta korostui tarve selkeälle dokumentaatiolle ja koulutukselle. Lähes kaikki vastaajat pitivät käyttöopasta ja kirjallista dokumentaatiota tärkeänä. Moni toivoi lisäksi käyttäjäkoulutusta (etänä tai paikan päällä) sekä lyhyitä perehdytysvideoita. Osa vastaajista nosti esiin itseopiskeltavan demon, jossa työkalua voisi kokeilla turvallisessa, tuotannosta erillisessä ympäristössä, sekä selkeästi nimettyä tukihenkilöä ongelmatilanteita varten. Nämä toiveet heijastavat tarvetta hallitulle ja vaiheistetulle käyttöönotolle, jossa käyttäjät voivat omaksua uuden työkalun rauhassa ja saada tukea tarvittaessa.

4.3.6 Luottamus, huolenaiheet ja riskit

Kyselyssä tarkasteltiin myös tekijöitä, jotka lisäävät luottamusta tekoälyn tuottamaan dataan, sekä huolia ja riskejä, joita vastaajat liittävät työkalun käyttöönottoon. Käyttäjien vastaukset luottamukseen liittyvään kysymykseen esitetty kuvassa 9 (Kuva 9).



Kuva 9. Kyselytutkimuksen vastaukset tekoälyyn luottamiseen liittyen. Lähteiden selkeää esittämistä ja ihmisen tekemää lopputarkastusta pidetään tärkeänä.

Luottamusta lisäävinä tekijöinä esiintyivät:

- **Mahdollisuus tarkistaa ja hyväksyä muutokset itse ennen tallennusta:** Tämä valittiin käytännössä kaikkien vastaajien toimesta, mikä osoittaa, että ihmisen lopullinen päätösvalta on keskeinen edellytys työkalun hyväksyttävyydelle.

- **Läpinäkyvyys ehdotusten syntytavasta:** Suuri osa vastaajista piti tärkeänä, että rikastusehdotusten perustelut ja tietolähteet ovat näkyvissä.
- **Käyttäjäpalautteen hyödyntäminen mallin parantamiseen:** Osa vastaajista koki tärkeäksi, että järjestelmä oppii palautteesta ja että käyttäjien kokemuksilla on näkyvä vaikutus mallin kehittymiseen.
- **Laadun mittarit ja vastuuroolit:** Muutamissa vastauksissa mainittiin erilaiset laadun mittarit (esim. osumatarkkuus) sekä vastuuhenkilön tai datavastaavan rooli ennen lopullista tallennusta.

Huolia ja riskejä koskevissa avoimissa vastauksissa nousi esiin erityisesti neljä teemaa:

- **Datan lähteisiin liittyvät rajoitteet:** Tiettyjen valmistajien data koettiin puutteelliseksi tai vaikeasti saatavaksi, mikä voi rajoittaa AI-työkalun mahdollisuuksia tuottaa luotettavia rikastusehdotuksia.
- **Virheellinen tai harhaanjohtava data:** Vastaajat pelkäsivät, että nimikkeisiin voi päätyä virheellisiä tai puutteellisia teknisiä tietoja, jos tekoälyn ehdotuksia ei tarkisteta riittävän huolellisesti.
- **Liiallinen luottamus tekoälyyn ja osaamisen katoaminen:** Useat vastaajat toivat esiin riskin, että työkalun toimiessa hyvin siihen voidaan alkaa luottaa liikaa, jolloin oma kriittinen arviointi ja osaaminen heikkenevät. Vastaavasti, jos työkalussa on ongelmia, virheiden määrä voi kasvaa.
- **Tietoturva ja integraatiot:** Osa huolista liittyi tietoturvaan, integraatioihin eri toimittajien järjestelmiin ja siihen, miten ulkoisista lähteistä haettava data käsitellään ja tallennetaan.

Monessa vastauksessa korostettiin, että vaikka tekoälytyökalu voi merkittävästi helpottaa työtä, lopullinen vastuu datan oikeellisuudesta tulee säilyä ihmisellä. Tämän tueksi työkalun tulee mahdollistaa vaivaton tarkistus ja hyväksyntä sekä tarjota riittävästi läpinäkyvyyttä ja kontekstia rikastusehdotuksille.

4.3.7 Yhteenveto kyselytutkimuksen keskeisistä havainnoista

Kyselytutkimuksen perusteella master datan käyttäjät ovat varsin yhtenäisiä siinä, että nykyinen nimikekirjaston käyttö on työlästä, hajanaista ja osin epäyhtenäisten käytäntöjen rasittamaa. Samanaikaisesti vastaajat tunnistavat tekoälyavusteisen rikastustyökalun tarjoamat mahdollisuudet merkittävinä. Työkalun odotetaan vähentävän manuaalista työtä, parantavan hinnoittelun ja komponenttivalinnan laatua, nopeuttavan rakenteiden käsittelyä sekä tukevan yhtenäisempiä nimike- ja laskentakäytäntöjä.

Tulokset kuitenkin korostavat, että toteutuksessa on huomioitava käyttäjien nostamat ehdot, joista keskeisimpiä ovat läpinäkyvät perustelut, mahdollisuus muutosten manuaaliseen hyväksyntään, selkeä esitystapa sekä riittävä koulutus. Ilman näitä elementtejä riskinä on liiallinen luottamus järjestelmään tai toisaalta sen käytön vastustus.

Kyselytutkimus tuotti sekä konkreettisia toiminnallisia vaatimuksia, kuten tulosten vertailunäkymän ja interaktiivisen muokkauksen, että laajemman näkemyksen käyttöönoton edellytyksistä. Tulokset korostivat erityisesti koulutuksen, selkeiden toimintatapojen ja hallintamallien merkitystä, jotta master datan rikastamiseen kehitetty AI-työkalu voidaan integroida osaksi kohdeyrityksen päivittäistä toimintaa kestävällä tavalla.

4.4 Projektin nykytilanne

Projektin nykytila kuvastaa vaihetta, jossa tekninen ratkaisu on kehitetty toimivaksi prototyypiksi, mutta sen laajamittainen käyttöönotto on vielä kesken. Tämä on tyypillinen tilanne Design Science Research -lähestymistavassa, jossa artefakti kehittyy iteratiivisesti ja sen arviointi tapahtuu vaiheittain (Hevner ja muut, 2004).

Alkuperäisessä suunnitelmassa hanke jaettiin kolmeen selkeään vaiheeseen: datan siivoukseen ja standardointiin, datan rikastamiseen sekä tekoälyn integrointiin osaksi

päivittäisiä työnkulkuja. Käytännössä nämä vaiheet eivät kuitenkaan toteutuneet lineaarisesti, vaan etenivät osittain limittäin ja jäivät joiltakin osin keskeneräisiksi. Tämä korostaa tekoälyhankkeille tyypillistä iteratiivista luonnetta sekä sitä, että datan laatu, prosessit ja tekninen toteutus ovat vahvasti keskinäisriippuvaisia.

Tässä työssä kehitetty tekoälyratkaisu on edennyt teknisesti toimivaan Proof-of-Concept- ja osittain MVP-tasoon, mutta sen täysi hyödyntäminen on riippuvainen erityisesti lähtödatan laadusta ja organisatoristen prosessien kypsyydestä. Näin ollen projektin nykytilaa voidaan kuvata siirtymävaiheena teknisestä validoinnista kohti operatiivista käyttöönottoa.

Keskeinen havainto on, että tekninen ratkaisu itsessään ei muodostu käyttöönoton pullonkaulaksi, vaan merkittävimmät haasteet liittyvät datan laatuun, yhtenäisyyteen ja hallintamalliin. Tämä tukee aiempaa tutkimusta, jonka mukaan tekoälyratkaisujen arvo realisoituu vasta, kun dataperusta ja prosessit ovat riittävän kypsiä (Ribeiro, 2024).

4.4.1 Web-sovelluksen käyttöönotto ja nykyinen toiminnallisuus

Master data -tekoälytyökalu on toteutettu web-sovelluksena, mutta sitä ei ole vielä otettu työntekijöiden käyttöön. Työkalu on tällä hetkellä testivaiheessa, ja sitä on arvioitu teknisestä ja toiminnallisesta näkökulmasta rajatussa kehitysympäristössä.

Web-sovellus mahdollistaa nimikkeiden haun tietokannasta sekä keskeisten tietokenttien tarkastelun. Tässä vaiheessa sen rooli on ensisijaisesti toimia analyysi- ja validointityökaluna, ei vielä operatiivisena rikastusratkaisuna.

Korvaavien tuotteiden automaattinen tunnistaminen ei ole vielä mahdollista. Tämä edellyttää riittävän kattavia ja yhdenmukaisesti tallennettuja teknisiä attribuutteja, kuten mittoja, sähköarvoja ja hyväksyntätietoja, joita ei nykyisessä tietokannassa ole systemaattisesti saatavilla.

Työkalu hyödyntää olemassa olevia komponenttitietoja sekä täydentää niitä ulkoisista lähteistä. Ensisijaisena lähteenä toimii EPLAN API ja toissijaisena Sähkönumerot.fi. Lisäksi kehitysvaiheessa on hyödynnetty rajattua verkkohakua (Bing Search) tilanteissa, joissa tietoa ei ole löytynyt ensisijaisista lähteistä. Tätä voidaan kuitenkin pitää kokeellisena ratkaisuna, eikä se sellaisenaan sovellu suoraan tuotantokäyttöön ilman tarkempaa lähdehallintaa ja validointimekanismeja.

4.4.2 Python-pohjainen rikastusprosessi ja tekninen toteutus

Koska lähtödatan laatu osoittautui puutteelliseksi, projektissa päädyttiin vaiheittaiseen lähestymistapaan, jossa perustason rikastus toteutettiin erillisillä Python-skripteillä ennen tekoälytyökalun laajempaa optimointia.

Skriptien avulla rikastettiin muun muassa seuraavia tietokenttiä:

- EPLAN-kategoriat
- Mitat
- Paino
- Alkuperämaa
- Tullikoodi
- Tekninen kuvaus
- Obsolete-status

Rikastusprosessi toteutettiin käsittelemällä aineisto rivitasolla, jossa nimike tunnistettiin MS BookPart -numeron perusteella ja sitä täydennettiin ulkoisista lähteistä.

Rikastus kohdistui ensisijaisesti keskeisiin laitevalmistajiin, jotka kattoivat noin 32 % aktiivisesta materiaalikirjastosta. Vaikka rikastus onnistui teknisesti näissä kohteissa, sen kokonaisvaikutus jäi rajalliseksi, koska rikastettu aineisto sijoittui pääosin jo ennestään laadukkaaseen osaan datasta.

4.4.3 Datan laadun haasteet

Projektissa havaittiin, että tekoälytyökalun tehokas toiminta edellyttää riittävän rikasta ja yhdenmukaista lähtödataa. Tilanne on kaksijakoinen, koska työkalun tarkoituksena on parantaa datan laatua, mutta se ei kykene toimimaan luotettavasti, mikäli lähtödatan taso on merkittävästi puutteellinen.

Duplikaattien tunnistaminen osoittautui erityisen haastavaksi. Automaattiset raportit tunnistivat duplikaatteina myös valmiita tuotteita, elinkaaren päässä olevia materiaaleja sekä nimikkeitä, joilla oli useita varastopaikkoja. Tämän vuoksi todellisten duplikaattien varmistaminen vaati manuaalista tarkastelua, ja analysoidussa aineistossa tunnistettiin lopulta noin 60 todellista duplikaattia.

Lisäksi havaittiin, että uutta duplikaattidataa syntyy jatkuvasti operatiivisessa toiminnassa, mikä korostaa tarvetta prosessien ja ohjeistusten kehittämiselle pelkän teknisen ratkaisun lisäksi.

Vastuullisuus- ja vaatimustenmukaisuustietojen (esim. CO_2, RoHS, REACH) rikastaminen osoittautui myös haastavaksi, eikä näitä tietoja kyetty luotettavasti täydentämään ulkoisista lähteistä.

4.4.4 Tietokannan siivous ja jatkokehitys

Seuraavassa kehitysvaiheessa painopiste siirtyy datan rakenteelliseen siivoukseen ja standardointiin. Tavoitteena on vähentää epäyhtenäisyyttä, poistaa epäolennaisia nimikkeitä sekä parantaa datan hyödynnettävyyttä rikastusprosessissa.

Projektissa kehitettiin myös uusia Material Group -luokkia EPLAN-luokittelun pohjalta. Näitä ei kuitenkaan vielä otettu käyttöön SAP-järjestelmässä, ja niiden käyttöönoton arvioitiin edellyttävän merkittäviä prosessimuutoksia sekä laajamittaista uudelleenluokittelua, mikä tekee toteutuksesta työlään.

Hintatiedon rikastamiseen liittyen havaittiin, että toimittajakohtaiset hinnat vaihtelevat merkittävästi, mikä edellyttää selkeää mallia hintojen hallintaan ennen automaation laajentamista.

4.4.5 Rikastustyökalun validointi ja suorituskyky

Projektin loppuvaiheessa toteutettiin rajattu demonstraatio, jossa rikastustyökalua testattiin 22 satunnaisesti valitulla nimikkeellä, joilta puuttui valmistajatieto.

Tulokset osoittivat, että:

- 15 tapauksessa rikastus oli täysin onnistunut
- 1 tapauksessa tulos oli täysin virheellinen
- Lopuissa tapauksissa tulokset olivat osittain oikeita tai epävarmoja

Tulosten perusteella voidaan todeta, että vaikka rikastustyökalu toimii teknisesti ja tuottaa käyttökelpoisia ehdotuksia, sen luotettavuus ei vielä riitä täysin automatisoituun käyttöön. Kaikki rikastusehdotukset vaativat edelleen manuaalisen tarkistuksen ja validoinnin ennen järjestelmään viemistä.

4.5 Tulosten yhteenveto ja synteesi

Tämän luvun tulokset osoittavat, että Arnon Oy:n master datan kehittämisen keskeisin haaste ei liity yksittäiseen teknologiaan, vaan datan laatuun, rakenteeseen ja hallintaan. Nykytila-analyysin perusteella data on laaja ja operatiivisesti merkittävä, mutta sen yhdenmukaisuus, täydellisyys ja ajantasaisuus eivät vielä kaikilta osin täytä tekoälypohjaisen rikastamisen edellytyksiä. Erityisesti toimittajanimien vaihtelu, puuttuvat attribuutit sekä havaitut duplikaattitapaukset osoittavat, että datan harmonisointi on välttämätön edellytys luotettavalle automaatiolle. Tämä havainto on linjassa datan laatua käsittelevän kirjallisuuden kanssa, jossa korostuvat tarkkuus, täydellisyys, yhdenmukaisuus ja ajantasaisuus keskeisinä laatudimensioina (Wang &

Strong, 1996; Batini & Scannapieco, 2016). Lisäksi useat tutkimukset osoittavat, että puutteellinen data heikentää merkittävästi analytiikan ja tekoälyratkaisujen suorituskykyä sekä luotettavuutta (Culot ja muut, 2024; Ribeiro, 2024).

Kyselytutkimus täydensi teknistä analyysiä käyttäjänäkökulmalla ja toi esiin datan laadun vaikutukset päivittäiseen työhön. Tulosten perusteella nimikekirjaston käyttäjät kokevat nykyiset työvaiheet työläiksi, mikä on tyypillinen ilmiö organisaatioissa, joissa master datan hallinta ei ole täysin standardoitua (Cao & Iansiti, 2021). Samanaikaisesti vastaajat suhtautuvat myönteisesti tekoälyavusteiseen rikastamiseen, kunhan ratkaisu säilyy läpinäkyvänä, käyttäjän hallittavana ja luottamusta rakentavana. Erityisesti mahdollisuus tarkistaa ehdotukset ennen tallennusta, nähdä käytetyt tietolähteet sekä tehdä muutoksia itse nousi keskeiseksi hyväksyttävyyden edellytykseksi. Tämä tukee aiempaa tutkimusta, jonka mukaan tekoälyratkaisujen käyttöönotto organisaatioissa edellyttää “human-in-the-loop” -periaatetta, jossa käyttäjä säilyttää lopullisen päätösvallan ja vastuun (Lazaros ja muut, 2026).

Projektin tekninen toteutus osoittaa, että Retrieval-Augmented Generation (RAG) -pohjainen lähestymistapa on perusteltu valinta tilanteessa, jossa tarvitaan sekä organisaation sisäistä tietoa että ulkoisia tietolähteitä rikastusehdotusten muodostamiseen. RAG-arkkitehtuuri mahdollistaa generatiivisen mallin yhdistämisen ajantasaiseen ja kontekstuaaliseen tietoon, mikä parantaa tuotettujen vastausten faktuaalista tarkkuutta ja vähentää hallusinaatioita (Lewis ja muut, 2020). Samalla tulokset kuitenkin osoittavat, että järjestelmän lopullinen hyöty realisoituu vasta, kun lähtödata on riittävän laadukasta ja prosessit on määritelty selkeästi. Tämä on linjassa tutkimusten kanssa, joissa korostetaan, että tekoälyjärjestelmien suorituskyky on suoraan riippuvainen käytetyn datan laadusta (“garbage in, garbage out”) (Hiniduma ja muut, 2024).

Projektin nykyinen kehitysvaihe on siten luonteeltaan valmistava. Sen keskeinen arvo ei ole vielä täysimittaisessa tuotantokäytössä, vaan siinä, että se on tuottanut teknisen

perustan sekä tunnistanut kriittiset datalliset ja organisatoriset edellytykset onnistuneelle käyttöönotolle. Tällaisia edellytyksiä ovat erityisesti datan harmonisointi, selkeät omistajuudet ja vastuut, yhtenäiset nimeämiskäytännöt sekä hallittu integraatio SAP-ympäristöön. Nämä havainnot tukevat Data Quality Management -kirjallisuutta, jossa datan laadun kehittämisen nähdään olevan jatkuva organisatorinen prosessi, ei pelkästään tekninen toimenpide (Batini & Scannapieco, 2016; Ibrahim ja muut, 2021).

Kokonaisuutena tutkimuksen tulokset muodostavat johdonmukaisen ja toisiaan tukevan kokonaisuuden, jossa tekninen analyysi ja käyttäjälähtöinen näkökulma täydentävät toisiaan. Tulosten synteesi osoittaa, että tekoälyavusteinen master datan rikastaminen on Arnonin toimintaympäristössä sekä mahdollinen että tarpeellinen ratkaisu, mutta sen onnistuminen edellyttää kolmen tekijän samanaikaista toteutumista. Ensinnäkin datan tulee olla riittävän laadukasta ja yhdenmukaista, jotta tekoälymallit voivat tuottaa luotettavia tuloksia. Toiseksi teknisen arkkitehtuurin tulee olla hallittu ja läpinäkyvä, jotta järjestelmä toimii ennustettavasti ja turvallisesti. Kolmanneksi käyttöönoton tulee olla käyttäjälähtöinen, jolloin ratkaisu tukee asiantuntijatyötä eikä korvaa sitä. Tämä kolmijako on linjassa myös tietojärjestelmätutkimuksen kanssa, jossa onnistuneet järjestelmät nähdään teknisten, organisatoristen ja inhimillisten tekijöiden yhteisvaikutuksena (Hevner ja muut, 2004).

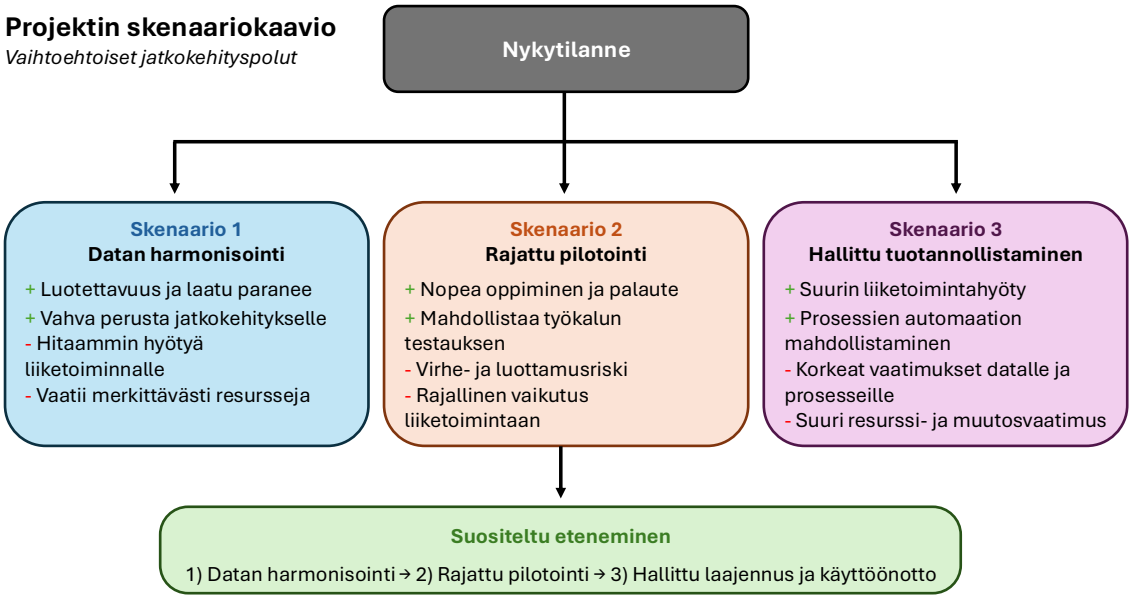
Edellä esitetyt havainnot osoittavat, että projektin toteutus ei vastannut täysin alkuperäistä vaiheistettua suunnitelmaa, vaan kehitys eteni iteratiivisesti ja osittain kokeiluluonteisesti. Tämä vahvistaa käsitystä siitä, että tekoälypohjaiset datanhallintaratkaisut eivät ole pelkästään teknisiä implementaatioita, vaan niiden onnistuminen edellyttää samanaikaista kehitystä datan laadussa, prosesseissa ja organisatorisessa ohjauksessa. Näin ollen projektin keskeinen kontribuutio ei ole pelkästään kehitetty artefakti, vaan myös ymmärrys niistä käytännön reunaehdoista, jotka määrittävät tekoälyratkaisun todellisen hyödyn organisaatiossa.

Näin muodostettu kokonaiskuva toimii perustana seuraaville luvuille, joissa tarkastellaan tarkemmin artefaktin arviointia, tutkimuksen rajoitteita sekä jatkokehityksen suuntaviivoja.

4.6 Skenaariopohjainen analyysi projektin jatkokehityksestä

Koska tekoälytyökalu ei ole vielä tuotantokäytössä ja artefaktin lopullinen validointi on kesken, projektin jatkon arviointi skenaariopohjaisesti on tässä vaiheessa metodologisesti perusteltu lähestymistapa. Skenaariopohjainen analyysi soveltuu erityisesti tilanteisiin, joissa toimintaympäristö on epävarma, muuttuva ja osin ennakoimaton, eikä luotettavaa empiiristä näyttöä kaikista vaikutuksista ole vielä saatavilla. Menetelmän keskeinen vahvuus on, että se mahdollistaa useiden vaihtoehtoisten kehityspolkujen systemaattisen tarkastelun yhden deterministisen ennusteen sijaan ja auttaa arvioimaan päätösten seurauksia erilaisissa olosuhteissa (Bradfield ja muut, 2005). Tällainen lähestymistapa tukee päätöksentekoa erityisesti silloin, kun teknologian käyttöönottoon liittyy organisatorisia, datallisia ja prosessuaalisia epävarmuuksia.

Design Science Research (DSR) -viitekehyksen näkökulmasta skenaariot toimivat tässä työssä artefaktin arvioinnin jatkeena. Hevnerin ym. (2004) mukaan DSR-tutkimuksen keskiössä on artefaktin rakentaminen, soveltaminen ja arviointi suhteessa tunnistettuun ongelmaan. Kun tuotantotason mittareita ei ole vielä saatavilla, skenaariopohjainen arviointi tarjoaa perustellun tavan analysoida, millaisia vaikutuksia eri kehitysvaihtoehdot todennäköisesti tuottavat organisaation toimintaan, datan laatuun ja käyttäjien työhön. Skenaariot tukevat design-syklin arviointivaihetta ja täydentävät empiirisen validoinnin puutetta teoreettisesti perustellulla analyysilla. Projektin skenaariokaavio esitetty kuvassa 10 (Kuva 10).



Kuva 10. Projektin skenaariokaavio. Kolme mahdollista jatkokehityspolkua projektin etenemiselle sekä suositeltu etenemismalli.

Skenaariokaavio havainnollistaa projektin kolme vaihtoehtoista jatkokehityspolkua. Ensimmäisessä skenaariossa painotetaan datan harmonisointia, toisessa rajattua pilotointia ja kolmannessa hallittua tuotannollistamista. Tässä analyysissä päädyttiin siihen tulokseen, että tarkoituksenmukaisin etenemismalli on vaiheittainen yhdistelmä, jossa datan laatu, käyttäjäpalaute ja tekninen käyttöönotto kehittyvät rinnakkain.

4.6.1 Datan harmonisointia painottava skenaario

Tässä skenaariossa projektin jatkokehitys keskittyy ensisijaisesti master datan laadun parantamiseen ennen laajempaa käyttöönottoa. Lähtökohtana on, että nykyinen data sisältää edelleen merkittäviä epäyhtenäisyyksiä, puutteita ja mahdollisia duplikaatteja, jotka heikentävät tekoälyratkaisun kykyä tuottaa luotettavia rikastusehdotuksia. Tämän vuoksi kehitystyö kohdistuisi nimikkeiden standardointiin, toimittajanimien harmonisointiin, puuttuvien attribuuttien täydentämiseen sekä epäselvien tai vanhentuneiden tietueiden siivoukseen.

Skenaarion vahvuus liittyy sen tuottamaan luotettavuuteen ja tekniseen vakauteen. Useat tutkimukset korostavat, että tekoälyjärjestelmien suorituskyky riippuu vahvasti käytetyn datan laadusta. Erityisesti epäyhtenäinen tai puutteellinen data voi heikentää mallien ennustettavuutta ja luotettavuutta (Batini & Scannapieco, 2016). Kun lähtödata saadaan yhtenäisemmäksi, RAG-pohjaisen järjestelmän hakutulosten relevanssi ja rikastusehdotusten tarkkuus paranevat. Tämä puolestaan vahvistaa käyttäjien luottamusta järjestelmään, mikä on keskeinen edellytys tekoälyratkaisujen hyväksynnälle organisaatioissa (Papenmeier ja muut, 2019; Bach ja muut, 2023).

Skenaarion suurin heikkous on sen hitaampi vaikutus näkyviin hyötyihin. Lyhyellä aikavälillä käyttäjien kokema työn tehostuminen voi jäädä vähäiseksi, koska painopiste on datan korjaamisessa eikä automaation laajentamisessa. Tästä huolimatta skenaario luo vahvan perustan myöhemmälle käyttöönotolle. Onnistumista voidaan arvioida esimerkiksi datan täydellisyysasteella, duplikaattien määrän vähenemisellä, nimeämiskäytäntöjen yhdenmukaisuudella sekä rikastusehdotusten hyväksymisprosentilla ilman merkittäviä korjauksia.

4.6.2 Rajattua pilotointia painottava skenaario

Toisessa skenaariossa ratkaisu viedään nopeasti rajatun käyttäjäryhmän testattavaksi nykyisellä datalla, vaikka data ei vielä olisi täysin harmonisoitu. Tavoitteena on kerätä käytännön palautetta käyttöliittymästä, rikastusehdotusten hyödyllisyydestä ja käyttöprosessin sujuvuudesta. Tämä lähestymistapa vastaa ketterän kehityksen periaatteita, joissa järjestelmää kehitetään iteratiivisesti käyttäjäpalautteen perusteella (Hevner ja muut, 2004).

Pilotoinnin keskeinen etu on nopea oppiminen. Käyttäjien palaute auttaa tunnistamaan, mitkä toiminnallisuudet tuottavat todellista arvoa, missä kohdin käyttöliittymä on epäselvä ja millaisia tietolähteitä pidetään luotettavina. Samalla voidaan arvioida käyttäjien luottamusta järjestelmään ja sitä, kuinka paljon manuaalista validointia tarvitaan ennen tietojen hyväksymistä. Tämä on tärkeää, sillä tutkimukset osoittavat,

että tekoälyjärjestelmien hyväksyttävyys riippuu merkittävästi läpinäkyvyydestä, kontrolloitavuudesta ja käyttäjän roolista päätöksenteossa (Shin, 2021; Sullivan ja muut, 2025).

Skenaarion keskeinen riski liittyy datan laatuun. Mikäli järjestelmä tuottaa heikkolaatuiseen dataan perustuvia virheellisiä tai epäluotettavia ehdotuksia, käyttäjien luottamus voi heikentyä nopeasti. Tämän vuoksi pilotointi edellyttää selkeää rajausstrategiaa, palautemekanismia sekä käyttäjien aktiivista tukea. Pilotin onnistumista voidaan arvioida useilla mittareilla, kuten käyttöasteella, hyväksyttyjen ehdotusten osuudella, käyttäjien tekemien korjausten määrällä sekä käyttäjätyytyväisyydellä. Aiempi tutkimus korostaa, että tekoälyjärjestelmien arvioinnissa teknisen suorituskyvyn lisäksi myös käyttäjäkokemus, luottamus järjestelmään ja käytännön hyödyllisyys ovat keskeisiä onnistumisen mittareita (Afroogh ja muut, 2024; Mazuruse ja muut, 2026).

4.6.3 Hallittua tuotannollistamista painottava skenaario

Kolmannessa skenaariossa organisaatio siirtyy asteittain kohti ratkaisun hallittua tuotannollistamista. Tämä edellyttää selkeää datan omistajuutta, määriteltyjä käyttöprosesseja, hyväksymiskäytäntöjä sekä teknistä integraatiota osaksi SAP-ympäristöä. Tällöin tekoälytyökalu muuttuu kokeilusta osaksi organisaation operatiivista tiedonhallintaa.

Tämän skenaarion hyöty on merkittävin vaikutus organisaation tehokkuuteen. Kun data, prosessit ja käyttöliittymä toimivat yhtenä kokonaisuutena, manuaalinen tiedonhaku vähenee, päätöksenteko nopeutuu ja datan laatu paranee jatkuvasti. Samalla kuitenkin myös vaatimukset kasvavat. Tarvitaan vahvempaa datanhallintaa, selkeää vastuunjakoa, riittävää resursointia ja jatkuvaa laadun seurantaa. Ilman näitä elementtejä riski järjestelmän osittaisesta epäonnistumisesta kasvaa. Tietojärjestelmätutkimuksessa onkin todettu, että järjestelmien onnistunut käyttöönotto edellyttää teknisten, organisatoristen ja inhimillisten tekijöiden yhteensovittamista (Hevner ja muut, 2004).

Onnistumista voidaan arvioida esimerkiksi läpimenoaikojen lyhenemisellä, virheellisten tietueiden vähenemisellä, rikastusehdotusten hyväksymisasteella sekä integraation sujuvuudella SAP-prosesseihin. Koska skenaarion vaikutukset ulottuvat laajasti organisaatioon, sen toteutus edellyttää myös systemaattista muutoksenhallintaa. Uudempi tutkimus korostaa, että digitaalisten ja tekoälyyn liittyvien muutosten onnistuminen riippuu vahvasti henkilöstön osallistamisesta, viestinnästä, koulutuksesta ja johdon sitoutumisesta muutokseen (Ibrahim ja muut, 2025; Tavantzis & Feldt, 2024).

4.6.4 Suositeltu etenemispolku

Skenaarioiden vertailun perusteella tarkoituksenmukaisin etenemispolku on vaiheistettu yhdistelmämalli, jossa ensin parannetaan datan laatua, sen jälkeen toteutetaan rajattu pilotti ja lopuksi siirrytään hallittuun laajennukseen. Pelkkä pilotointi ilman datan harmonisointia tuottaisi todennäköisesti lyhyen aikavälin oppia, mutta ei ratkaisisi keskeisiä laatuongelmia. Vastaavasti pelkkä datan siivous ilman käyttäjätestausta voisi viivästyttää konkreettisen hyödyn syntymistä.

Vaiheistettu lähestymistapa yhdistää iteratiivisen kehityksen ja laadun systemaattisen parantamisen, mikä on linjassa sekä DSR-menetelmän että modernin tekoälykehityksen parhaiden käytäntöjen kanssa. Artefaktia kehitetään vaiheittain, sitä arvioidaan todellisessa käyttöympäristössä ja kehityspäätökset perustuvat sekä käyttäjäpalautteeseen että datan laadullisiin mittareihin. Skenaariopohjainen analyysi toimii näin välivaiheen arviointina tilanteessa, jossa lopullinen tuotantodataa koskeva näyttö puuttuu, mutta strategisia päätöksiä jatkokehityksestä on silti tehtävä.

4.6.5 Yhteys tutkimuskysymyksiin

Tulosten synteesi vastaa suoraan tutkimuksessa asetettuihin tutkimuskysymyksiin. Ensinnäkin tulokset osoittavat, että keskeisimmät master datan laatuun liittyvät haasteet liittyvät täydellisyyteen, yhdenmukaisuuteen ja ajantasaisuuteen, mikä vahvistaa

aiempien tutkimusten havaintoja datan laadun keskeisistä ulottuvuuksista (Wang & Strong, 1996).

Toiseksi havaittiin, että tekoälypohjaiset rikastusmenetelmät, erityisesti RAG-arkkitehtuuriin perustuvat ratkaisut, voivat merkittävästi tukea datan rikastamista ja eheyttämistä, mutta niiden tehokkuus on suoraan sidoksissa lähtödatan laatuun. Kolmanneksi tulokset osoittavat, että tekoälyratkaisu ei korvaa asiantuntijatyötä, vaan toimii sitä tukevana välineenä, mikä korostaa ihmisen roolia päätöksenteossa.

Lisäksi skenaariopohjainen analyysi osoittaa, että onnistunut käyttöönotto edellyttää vaiheittaista etenemistä, jossa yhdistyvät datan laadun parantaminen, käyttäjälähtöinen pilotointi ja hallittu tuotannollistaminen. Täten tutkimus tarjoaa kokonaisvaltaisen vastauksen siihen, miten master dataa voidaan kehittää tekoälyn avulla sekä teknisestä että organisatorisesta näkökulmasta.

5 Pohdinta

Tässä luvussa tutkimuksen tuloksia tarkastellaan suhteessa aiempaan kirjallisuuteen sekä arvioidaan niiden merkitystä sekä tutkimuksellisesta että käytännön näkökulmasta. Erityisesti tarkastellaan, miten saadut tulokset tukevat tai haastavat aiempia havaintoja datan laadun ja tekoälyratkaisujen välisestä suhteesta.

Tulokset vahvistavat vahvasti kirjallisuudessa esitettyä näkemystä siitä, että datan laatu on keskeisin yksittäinen tekijä tekoälyratkaisujen onnistumisessa (Ehrlinger ja muut, 2022; Hiniduma ja muut, 2024). Samalla tutkimus osoittaa myös, että pelkkä teknisesti toimiva ratkaisu ei riitä onnistuneeseen käyttöönottoon. Lisäksi tarvitaan organisatorista kypsyyttä, selkeitä prosesseja ja käyttäjien luottamusta järjestelmään.

5.1 Projektin tekniset ratkaisut

Azure AI Foundry toimi projektissa keskeisenä kehitysympäristönä, joka mahdollisti kielimalleihin perustuvien dataputkien, hakurakenteiden ja integraatioiden hallitun suunnittelun ja testauksen. Alusta tuki erityisesti Retrieval-Augmented Generation (RAG) -arkkitehtuurin toteuttamista, jossa generatiivinen kielimalli yhdistetään ulkoisista tietolähteistä haettuun kontekstiin (Lewis ja muut, 2020). Tällainen lähestymistapa on perusteltu teollisessa ympäristössä, jossa ajantasaisen ja lähdeankkuroidun tiedon hyödyntäminen on keskeistä.

Kehitysvaiheessa hyödynnettiin myös rajattua internet-hakua (Bing Search) täydentävänä tiedonlähteenä tilanteissa, joissa tietoa ei ollut saatavilla rajapinnoista, kuten EPLAN API:sta. Internet-haku osoittautui hyödylliseksi yksittäisissä tapauksissa, mutta siihen liittyy sekä kustannuksiin että tiedon luotettavuuteen liittyviä epävarmuuksia. Tämän vuoksi sen käyttö tuotantoympäristössä edellyttää tarkempaa kustannus–hyötyanalyysiä sekä lähdehallinnan ja validointimekanismien määrittelyä.

Alkuperäinen tavoite integroida ratkaisu Microsoft Teams -ympäristöön osoittautui teknisesti haastavaksi, minkä vuoksi lopullinen toteutus tehtiin erillisenä web-sovelluksena. Tämä ratkaisu paransi kehitystyön hallittavuutta ja testattavuutta, mutta siirsi osan integraatiohaasteista myöhempään vaiheeseen.

Projektin keskeinen tekninen havainto liittyy kehitysjärjestykseen. Tekoälyratkaisun kehittäminen ennen systemaattista datan profilointia ja siivousta johti tilanteeseen, jossa työkalun potentiaalia ei voitu hyödyntää täysimääräisesti. Tämä tukee datanhallinnan tutkimusta, jonka mukaan datan profilointi ja laadun arviointi tulisi tehdä ennen kehittyneiden analytiikka- ja tekoälyratkaisujen käyttöönottoa (Batini & Scannapieco, 2016).

Jatkokehityksen näkökulmasta keskeinen kysymys liittyy tietolähdestrategiaan. Mikäli rajapintapohjaiset lähteet, kuten EPLAN API, kattavat riittävän osan tarvittavista attribuuteista, voidaan rakentaa kustannustehokas ja hallittava rikastusratkaisu ilman jatkuvaa internet-hakua. Tällöin internet-pohjainen rikastus voidaan rajata kertaluonteisiin massarikastuksiin, kun taas jatkuva ylläpito perustuu luotettaviin ja kontrolloituihin rajapintoihin. Tällainen hybridimalli yhdistää joustavuuden ja hallittavuuden, mikä on keskeinen vaatimus teollisissa tietojärjestelmissä.

5.2 Projektista saadut opit ja retrospektiivinen analyysi

Projektin retrospektiivinen tarkastelu osoittaa, että keskeisin oppi liittyy datan ratkaisevaan merkitykseen tekoälyprojektin onnistumisessa. Käytännön toteutus vahvisti kirjallisuudessa laajasti tunnistetun periaatteen, jonka mukaan tekoälyjärjestelmien suorituskyky on suoraan riippuvainen lähtödatan laadusta (Ehrlinger ja muut, 2022; Hiniduma ja muut, 2024). Datan hajanaisuus, puutteellisuus ja epäyhtenäisyys rajoittivat merkittävästi järjestelmän kykyä tuottaa luotettavia rikastusehdotuksia, mikä korosti tarvetta systemaattiselle datan siivoukselle ennen automaation laajentamista.

Toinen keskeinen oppi liittyy projektinhallintaan ja omistajuuteen. Retrospektiiviset havainnot osoittavat, että kehitystyö eteni paikoin ilman selkeää vastuunjakoa ja yhtenäistä ohjausmallia. Tämä johti tilanteisiin, joissa kehityssuunta vaihteli ja osa ratkaisuista jäi irrallisiksi kokeiluiksi. Tietojärjestelmätutkimuksessa onkin korostettu, että teknologiaprojektien onnistuminen edellyttää vahvaa hallintamallia, selkeää päätöksentekorakennetta ja riittävää resursointia (Hevner ja muut, 2004).

Kolmas keskeinen havainto liittyy organisaation oppimiseen. Projektin aikana ymmärrys tekoälyn mahdollisuuksista ja rajoitteista syveni merkittävästi, ja alkuperäiset odotukset muuttuivat realistisemmiksi. Kävi selväksi, että tekoäly ei korvaa asiantuntijaa, vaan toimii hänen työnsä tukena. Tämä havainto on linjassa tutkimuksen kanssa, jossa korostetaan “human-in-the-loop” -periaatetta tekoälyjärjestelmien käyttöönotossa (Lazaros ja muut, 2026).

5.3 Osaamisen rappeutuminen AI-käytössä

Tekoälyratkaisujen käyttöönottoon liittyy riski osaamisen heikkenemisestä, jota tutkimuskirjallisuudessa kutsutaan osaamisen rappeutumiseksi (*skill erosion* tai *deskilling*). Viimeaikainen tutkimus osoittaa, että kognitiivinen automaatio ja tekoälyn jatkuva käyttö voivat heikentää työntekijöiden taitoja, jos omia valmiuksia ei enää harjoiteta aktiivisesti ja järjestelmän tuottamiin vastauksiin luotetaan liikaa (Rinta-Kahila ja muut, 2023). Käytännössä tämä voi tarkoittaa esimerkiksi sitä, että tiedonhaku, luokittelu tai teknisten tietojen vertailu siirtyy kokonaan tekoälylle, jolloin asiantuntijan oma analysointikyky heikkenee ajan myötä. Tutkijoiden mukaan ilmiö voi muodostua organisaatioille merkittäväksi riskiksi, jos työntekijät eivät enää kykene suorittamaan tehtäviään ilman automaation tukea (Athow, 2025). Aiempi tutkimus myös osoittaa, että tekoälyn käyttö analyysissä ja päätöksenteossa voi johtaa osaamisen rappeutumiseen sekä asiantuntijoiden oman arviointikyvyn heikkenemiseen, mikäli ihmiset alkavat nojata järjestelmän tuottamiin ehdotuksiin ilman aktiivista kriittistä arviointia (Natali ja muut, 2025).

Toinen keskeinen ilmiö on *automation bias*, jossa käyttäjä hyväksyy järjestelmän tuottaman tuloksen ilman riittävää kriittistä arviointia (Rosbach ja muut, 2024). Master data -ympäristössä tämä voisi tarkoittaa esimerkiksi sitä, että tekoälyn ehdottama luokitus tai tekninen tieto hyväksytään tarkistamatta sen oikeellisuutta. Pitkällä aikavälillä tällainen toimintatapa voi heikentää asiantuntijoiden arviointikykyä ja lisätä virheriskiä. Lisäksi ihmisen ja koneen yhteistyötä käsittelevä tutkimus osoittaa, että automaatio voi vahvistaa käyttäjän liiallista luottamusta järjestelmään, mikä vähentää oman harkinnan aktiivista käyttöä (Vicente & Matute, 2026).

Näiden riskien vuoksi tekoälytyökalun suunnittelussa tulisi tukea käyttäjän oppimista eikä korvata sitä. Työkalun tulisi toimia päätöksenteon tukena, ei itsenäisenä päätöksentekijänä. Käyttäjän on nähtävä, mihin ehdotus perustuu ja säilytettävä vastuu lopullisesta hyväksynnästä. Tätä noudattamalla voidaan vähentää liiallista riippuvuutta järjestelmästä ja ylläpitää organisaation osaamistasoa.

6 Johtopäätökset

Tässä luvussa esitetään tutkimuksen keskeiset johtopäätökset suhteessa asetettuihin tutkimuskysymyksiin. Lisäksi tarkastellaan tutkimuksen käytännön merkitystä kohdeorganisaatiolle, sen akateemista kontribuutiota sekä jatkokehityksen suuntaviivoja.

6.1 Keskeiset johtopäätökset tutkimuskysymyksittäin

Tutkimuskysymys 1: Mitkä master datan attribuutit ovat kriittisimpiä tilausprosessin onnistumiselle?

Master datan kriittisimmät attribuutit liittyvät täydellisyyteen, yhdenmukaisuuteen, ajantasaisuuteen ja tarkkuuteen. Erityisesti tuotekoodit, valmistajatiedot, tekniset attribuutit sekä kaupalliset tiedot osoittautuivat keskeisiksi tilausprosessin kannalta. Puutteellinen data lisää suoraan manuaalista työtä ja virheriskiä, mikä on linjassa datan laatututkimuksen kanssa (Wang & Strong, 1996).

Tutkimuskysymys 2: Millä menetelmillä master dataa voidaan eheyttää ja rikastaa automaattisesti?

Tutkimuksen perusteella tehokkain lähestymistapa master datan eheyttämiseen ja rikastamiseen on yhdistelmä useista menetelmistä. Näihin kuuluvat datan harmonisointi, sääntöpohjainen validointi, ulkoisten tietolähteiden hyödyntäminen sekä tekoälypohjainen rikastus.

Erityisesti Retrieval-Augmented Generation (RAG) -arkkitehtuuri osoittautui soveltuvaksi ratkaisuksi, koska se mahdollistaa generatiivisen kielimallin yhdistämisen ajantasaiseen ja kontekstuaaliseen tietoon. Käytännön toteutus kuitenkin osoitti, että rikastus kohdistui pääosin jo ennestään laadukkaaseen dataan, mikä rajoitti kokonaisvaikutusta.

Tutkimuskysymys 3: Kuinka kaupallinen AI-rikastustyökalu suoriutuu verrattuna manuaalisiin menetelmiin master datan rikastuksessa?

Tutkimus osoittaa, että tekoälypohjainen rikastustyökalu voi vähentää manuaalista työtä, mutta ei korvaa sitä. Validointitulokset (15/22 onnistui täysin) osoittavat, että kaikki rikastukset vaativat edelleen manuaalisen tarkistuksen. Näin ollen työ siirtyy tiedonhausta validointiin.

Tutkimuskysymys 4: Miten AI-ratkaisut täydentävät perinteistä rikastustyötä?

Tulosten perusteella tekoäly toimii ensisijaisesti asiantuntijatyötä tukevana välineenä. Se voi nopeuttaa tiedonhakua, ehdottaa rikastuksia ja yhdistää tietoa eri lähteistä, mutta ei korvaa asiantuntijan roolia päätöksenteossa.

Kyselytutkimus osoitti, että käyttäjät arvostavat erityisesti läpinäkyvyyttä, lähdeviitteitä ja mahdollisuutta tarkistaa ehdotukset ennen hyväksyntää. Tämä korostaa human-in-the-loop -periaatteen merkitystä tekoälyratkaisujen käyttöönotossa (Lazaros ja muut, 2026).

Tekoälyn rooli voidaan siten nähdä “älykkäänä avustajana”, joka parantaa työn tehokkuutta, mutta säilyttää päätöksenteon ihmisellä.

Tutkimuskysymys 5: Mitä muutoksia tarvitaan, jotta eheä ja rikastettu master data voidaan ylläpitää pysyvästi?

Tulosten perusteella pysyvä master datan laatu edellyttää teknisten ratkaisujen lisäksi organisatorisia muutoksia. Keskeisiä tekijöitä ovat selkeä datan omistajuus, yhtenäiset prosessit, jatkuva laadunvalvonta sekä käyttäjien koulutus.

Lisäksi tutkimuksessa tunnistettiin riski osaamisen rappeutumisesta, mikäli käyttäjät alkavat luottaa liikaa tekoälyyn ilman kriittistä arviointia. Tämän vuoksi on tärkeää säilyttää tasapaino automaation ja asiantuntijuuden välillä.

6.2 Käytännön merkitys kohdeyritykselle

Tutkimus tarjoaa Arnon Oy:lle konkreettisen kehityspolun master datan hallinnan parantamiseen. Keskeinen suositus on vaiheittainen eteneminen, jossa ensin parannetaan datan laatua, tämän jälkeen toteutetaan rajattu pilotointi ja lopuksi siirrytään hallittuun käyttöönottoon.

Tämä lähestymistapa minimoi riskit ja mahdollistaa jatkuvan oppimisen. Lisäksi tutkimus korostaa, että investoinnit tekoälyratkaisuihin ovat perusteltuja vain, mikäli datan laatuun ja prosesseihin panostetaan samanaikaisesti.

6.3 Akateeminen merkitys

Tutkimus tuottaa käytännönläheistä tietoa tekoälypohjaisten ratkaisujen soveltamisesta master datan hallintaan teollisessa kontekstissa. Se osoittaa, että RAG-pohjaiset ratkaisut ovat lupaava lähestymistapa, mutta niiden onnistuminen on vahvasti sidoksissa datan laatuun ja organisatorisiin tekijöihin.

Lisäksi tutkimus täydentää DSR-kirjallisuutta osoittamalla, miten artefaktin arviointia voidaan tehdä skenaariopohjaisesti tilanteessa, jossa täysimittaista käyttöönottoa ei ole vielä saavutettu.

6.4 Tutkimuksen rajoitteet ja jatkokehitys

Tutkimuksen keskeinen rajoite liittyy siihen, että tekoälytyökalu ei ole vielä tuotantokäytössä. Tämän vuoksi tulokset perustuvat nykytila-analyysiin, käyttäjäkyselyyn ja prototyyppivaiheen havaintoihin.

Jatkokehityksessä keskeistä on työkalun pilotointi, laajempi käyttöönotto sekä vaikutusten mittaaminen esimerkiksi tehokkuuden, virheiden määrän ja käyttäjätyytyväisyyden näkökulmasta.

6.5 Lopullinen arvio ja suositus

Tutkimuksen perusteella tekoälyavusteinen master datan rikastaminen on perusteltu ja potentiaalisesti merkittävä kehityssuunta Arnon Oy:lle. Sen onnistuminen edellyttää kuitenkin kokonaisvaltaista lähestymistapaa, jossa yhdistyvät laadukas data, toimiva tekninen ratkaisu ja käyttäjälähtöinen käyttöönotto.

Tekoäly ei korvaa asiantuntijatyötä, mutta oikein toteutettuna se voi merkittävästi tehostaa sitä ja parantaa datan laatua. Näin ollen suosituksena on edetä vaiheittain ja varmistaa, että jokainen kehitysvaihe tukee seuraavaa.

Lähteet

Afroogh, S., Akbari, A., Alambeigi, H., Kargar, M., & Malone, E. (2024). Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications.* https://doi.org/10.1057/s41599-025-04374-1

Agate, M. (2025). Artificial intelligence methods and approaches to improve data quality in healthcare data. *Intelligent Systems with Applications*. https://doi.org/10.1016/j.ailsci.2025.100135

Athow, D. (2025). Researchers warn that skill erosion caused by AI could have a devastating and lasting impact on businesses - but it may already be too late. Techradar. https://www.techradar.com/pro/researchers-warn-that-skill-erosion-caused-by-ai-could-have-a-devastating-and-lasting-impact-on-businesses-but-it-may-already-be-too-late

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2023). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human Computer Interaction 2022*. https://doi.org/10.48550/arXiv.2304.08795

Batini, C., & Scannapieco, M. (2016). *Data and information quality: Dimensions, principles and techniques*. Springer. https://doi.org/10.1007/978-3-319-24106-7

Bell, E., & Warren, V. (2023). Illuminating a methodological pathway for doctor of business administration researchers: Utilizing case studies and mixed methods for applied research. *Social Sciences & Humanities Open*. https://doi.org/10.1016/j.ssaho.2022.100391

Bradfield, R., Wright, G., Burt, G., Cairns, G., & van der Heijden, K. (2005). The origins and evolution of scenario techniques in long range business planning. *Futures, 37(8), 795–812*. https://doi.org/10.1016/j.futures.2005.01.003

Cao, L., & Iansiti, M. (2021). Data Governance, Interoperability, and Standardization: Organizational Adaptation to Privacy Regulations. *In 29th Annual Americas Conference on Information Systems: Diving into Uncharted Waters* (Panama City, August 10–12, 2023).

https://research.hhs.se/esploro/outputs/conferencePaper/Data-Governance-Interoperability-and-Standardization-Organizational/991001579399406056

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the 2016 International Conference on Management of Data*. https://doi.org/10.1145/2882903.2912574

Culot, G., Orzes, G., Sartor, M., & Nassimbeni, G. (2024). Artificial intelligence in supply chain management: A systematic literature review of empirical studies and research directions. *Computers in Industry*. https://doi.org/10.1016/j.compind.2024.104132

Dwivedi, Y. K., et al. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International *Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Ehrlinger, L. & Wöß, W. (2022). A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, 5. https://doi.org/10.3389/fdata.2022.850611

Gualo, F., Caballero, I., Rodríguez, M., & Piattini, M. (2023). A Data Quality Model for Master Data Repositories. *Informatica: An International Journal of Computing and Informatics*, 47(4), 603–620. https://doi.org/10.15388/23-INFOR534

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2), Article 4. https://aisel.aisnet.org/sjis/vol19/iss2/4

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. https://doi.org/10.2307/25148625

Hiniduma, K., Byna, S., Bez, J. L., & Madduri, R. (2024). *AI data readiness inspector (AIDRIN) for quantitative assessment of data readiness for AI*. arXiv. https://doi.org/10.48550/arXiv.2406.19256

Ibrahim, A., Mohamed, I., & Satar, N. S. M. (2021). Factors influencing master data quality: A Systematic Review. *International Journal of Advanced Computer Science and*

Applications, 12(2). https://thesai.org/Downloads/Volume12No2/Paper_24-Factors_Influencing_Master_Data_Quality.pdf

Ibrahim, A. K., Okafor, C. M., Wedraogo, L., Essandoh, S., Sakyi, J. K., Babalola, A. S., & Adenuga, M. A. (2025). Analysis of change management strategies during digital transformation projects. *International Journal of Finance and Management Research*, 6(1), 316–326. https://doi.org/10.54660/.IJFMR.2025.6.1.316-326

Jarrahi, M. H., Memariani, A., & Guha, S. (2024). *The Principles of Data-Centric AI (DCAI)*. arXiv. https://doi.org/10.48550/arXiv.2211.14611

Kim, F. (2025). *Poor data quality is a full-blown crisis: A 2024 customer insight report*. Datalere. https://datalere.com/articles/poor-data-quality-is-a-full-blown-crisis-a-2024-customer-insight-report

Lazaros, K., Vrahatis, A. G., & Kotsiantis, S. (2026). Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications. *Entropy*, 28(4), 377. https://doi.org/10.3390/e28040377

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv. https://doi.org/10.48550/arXiv.2005.11401

Matayo, S. (2026). *Why data governance fails in many organizations: The IT-business divide.* DATAVERSITY. https://www.dataversity.net/articles/why-data-governance-fails-in-many-organizations-the-it-business-divide/

Mazuruse, G., Nyagadza, B., Chifurira, R., Muvuti, A., & Matsiwira, L. (2026). Artificial intelligence (AI) adoption and satisfaction in management education research: An explanatory-predictive hybrid SEM-RF approach. *Strategic Business Research*, 2(1), 100102. https://doi.org/10.1016/j.sbr.2026.100102

McKinsey & Company. (2024). Master data management: The key to getting more from your data. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/master-data-management-the-key-to-getting-more-from-your-data

Microsoft. (2026). *Retrieval augmented generation (RAG)*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/foundry/concepts/retrieval-augmented-generation

Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 10160–10171). ELRA & ICCL. https://aclanthology.org/2024.lrec-main.887/

Nair, M. K. (2025). The strategic role of master data management in modern supply chain operations. *International Journal of Information Technology and Management Information Systems*, 16(2), 240–253. https://doi.org/10.34218/IJITMIS_16_02_017

Natali, C., Marconi, L., Dias Duran, L. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review*, 58, 356. https://doi.org/10.1007/s10462-025-11352-1

Pansara, R. (2023). Master data management in manufacturing industry. *International Journal of Scientific and Research Publications*, 13(11), 355. https://doi.org/10.29322/IJSRP.13.11.2023.p14335

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. arXiv. https://doi.org/10.48550/arXiv.1907.12652

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. https://www.researchgate.net/publication/284503626_A_design_science_research_methodology_for_information_systems_research

Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11), 1190–1201. https://doi.org/10.14778/3137628.3137631

Ribeiro, G. (2024, August 1). The critical role of data quality in AI. Forbes. https://www.forbes.com/councils/forbestechcouncil/2024/08/01/the-critical-role-of-data-quality-in-ai/

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The vicious circles of skill erosion: A case study of cognitive automation. *Journal of the Association for Information Systems*, 24(5), 1378–1412. https://doi.org/10.17705/1jais.00829

Rosbach, E., Ganz, J., Ammeling, J., Riener, A., & Aubreville, M. (2024). Automation bias in AI-assisted medical decision-making under time pressure in computational pathology. arXiv. https://doi.org/10.48550/arXiv.2411.00998

Rana, K., Goyal, P., & Sharma, G. (2024). Dual-branch convolutional neural network for robust camera model identification. *Expert Systems with Applications*, 246, 123456. https://doi.org/10.1016/j.eswa.2023.123456

SAP. (2025). *What is data quality?* SAP. https://www.sap.com/finland/resources/what-is-data-quality

Sargiotis, D. (2024). Data quality management: Ensuring accuracy and reliability. *In Data governance* (pp. 197–216). Springer Nature. https://doi.org/10.1007/978-3-031-67268-2_5

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance. *International Journal of Human-Computer Studies*, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Sullivan, V., & Weger, K. (2025). Transparency and explainability in AI-assisted decision making: Effects on trust, perceived reliability, confidence, and ease of understanding. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 69(1). https://doi.org/10.1177/10711813251369473

Tamm, H. C., & Nikiforova, A. (2024). From data quality for AI to AI for data quality: A systematic review of tools for AI-augmented data quality management in data warehouses. arXiv. https://doi.org/10.48550/arXiv.2406.10940

Tavantzis, T., & Feldt, R. (2024). Human-centered AI transformation: Exploring behavioral dynamics in software engineering. arXiv. https://doi.org/10.48550/arXiv.2411.08693

Vallepu, R. (2025). Enhancing AI and machine learning performance through effective master data management. *International Journal of Computer Applications*, 186(69). https://www.ijcaonline.org/archives/volume186/number69/vallepu-2025-ijca-924535.pdf

Vicente, L., & Matute, H. (2026). Warning people about the risk of AI error mitigates human acquisition of AI bias. *Cognitive Research: Principles and Implications*, 11(1), Article 36. https://doi.org/10.1186/s41235-026-00726-w

Vihavainen, K. (2014). *Managing data integrity as part of master data management* (Master's thesis, Helsinki Metropolia University of Applied Sciences). https://urn.fi/URN:NBN:fi:amk-201404305466

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Yang, X., Fu, Y., & Amin, M. (2025). The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*, 5, 323–350. https://doi.org/10.1007/s44230-025-00106-5

Liitteet

Liite 1. Kyselytutkimuksen runko

KYSELY: Tekoälyavusteinen datan rikastaminen

1. Mikä on roolisi organisaatiossa?
2. Kuinka usein työskentelet ostomateriaalien/nimikkeiden parissa?
3. Kuinka pitkään olet käyttänyt Falcony/Incy-työkalua nimikkeiden avaukseen?
4. Mihin käyttötarkoituksiin hyödynnät materiaalitietoja/nimikekirjastoa omissa työtehtävissäsi? (voit valita monta)
5. Minkä koet haastavimmaksi tai työläimmäksi nimikekirjaston parissa työskennellessä? (voit valita monta)
6. Mitkä materiaalitiedot (attribuutit) ovat mielestäsi olennaisia? (voit valita monta)
7. Tulisiko materiaalit/nimikkeet olla luokiteltuna kategorioihin ominaisuuksiensa mukaan?
8. Koetko nykyisen nimikkeiden rakenteen selkeäksi ja riittäväksi?
9. Kuinka hyödyllisenä pidät toimintoa, jossa työkalu ehdottaa tai täyttää automaattisesti komponentin/materiaalin tietoja tuotekoodin perusteella?
10. Kuinka hyödyllisenä pidät, että työkalua voisi kouluttaa käyttäjän tekemien korjausten perusteella?
11. Kuinka hyödyllisenä pidät, että työkalu pystyy perustelemaan rikastusehdotuksensa materiaalin tiedoista?
12. Millainen tulosten esitystapa olisi sinulle selkein? (voit valita monta)
13. Millainen koulutus tai tuki helpottaisi työkalun käyttöönottoa? (voit valita monta)
14. Mitkä asiat lisäisivät luottamusta tekoälyn tekemään datarikastukseen? (voit valita monta)
15. Mitä huolia tai riskejä liität tällaisen työkalun käyttöönottoon?
16. Miten tekoälyavusteinen datan rikastaminen voisi helpottaa omaa työtäsi?
17. Millaista lisäarvoa tällainen työkalu voisi tuoda yritykselle kokonaisuutena?
18. Onko sinulla muita ideoita tai toiveita työkalun toiminnallisuuksiin liittyen?