Tampereen yliopisto

Juuso Halonen

TIETOPANKKI APUNA DATATIETEELLISTEN PROJEKTIEN TOISTETTAVUUDESSA

Diplomityö
Johtamisen ja talouden tiedekunta
Tarkastaja: Henri Pirkkalainen
Tarkastaja: Marko Seppänen
Toukokuu 2023

TIIVISTELMÄ

Juuso Halonen: Tietopankki apuna datatieteellisten projektien toistettavuudessa
Diplomityö
Tampereen yliopisto
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
Toukokuu 2023

Datamäärien kasvu ja analytiikan monimutkaistuminen ovat vaikeuttaneet datatieteellisten projektien toistettavuutta. Huono toistettavuus hidastaa uusien projektien toteuttamista ja vaikeuttaa tiedon säilyttämistä ja oppimista aiemmista projekteista. Toistettavuus hankaloituu, kun projektin etenemisen aikana tehdyistä valinnoista ja päätöksistä ei ole koottu tietoa. Ongelma kiteytyy tiedon heikkoon jakamiseen ja uudelleenkäyttöön, minkä vuoksi tietämyksenhallinta on toistettavuushaasteen äärellä keskeistä. Tietopankit ovat yksi tietämyksenhallinnan keino ratkaista edellä mainittuja ongelmia. Toistettavuushaasteen ratkaisulla voidaan mahdollistaa myös datatieteellisiä projekteja toteuttavan yrityksen liiketoiminnan kasvu.

Tässä tutkimuksessa selvitettiin, millaisen tietopankin avulla datatieteellisten projektien toistettavuutta lisätään. Tutkimuksen toteutustapa oli laadullinen tutkimus. Tutkimus tehtiin tapaustutkimuksena datatieteellisiä projekteja toteuttavan yrityksen kontekstissa teemahaastattelujen avulla. Haastateltavat olivat yrityksen työntekijöitä sekä potentiaalisia tietopankin käyttäjiä.

Tutkimuksen perusteella havaittiin, että datatieteellisten projektien toistettavuuden mahdollistaa tietopankki, joka sisältää projektiin tarvittavia resursseja eri abstraktiotasoilla. Lisäksi tietopankkiin tulee toteuttaa käyttöä tukevia ominaisuuksia kuten hakutoiminto sekä versionhallinta, ja tietopankin käyttö tulee olla huomioitu datatieteellisessä prosessimallissa.

Avainsanat: Tietämyksenhallinta, Datatieteelliset projektit, Tietopankki, Toistettavuus

ABSTRACT

Juuso Halonen: Knowledge repository for supporting the reproducibility of data science projects
Master's Thesis
Tampere University
Master's degree programme in information and knowledge management
May 2023

The increase in data volumes and the complexity of analytics have made the reproducibility of data science projects difficult. Poor level of reproducibility slows down the implementation of new projects and hinders the retention of knowledge and makes it harder for organizations to learn from past projects. Reproducibility becomes difficult when knowledge about the choices and decisions made during the project has not been systematically gathered. The problem stems from inadequate sharing and reuse of knowledge, making knowledge management crucial in addressing the challenge of reproducibility. Knowledge repositories are one means of knowledge management to solve the above-mentioned problems. A solution to the reproducibility challenge can also enable the business growth of a company implementing data science projects.

This study aims to find out what kind of knowledge repository increases the reproducibility of data science projects. The research was qualitative, and it was conducted as a case study in the context of a company doing data science projects by means of thematic interviews. The interviewees were employees of the company and potential users of the knowledge repository.

Based on the research, it was found that the reproducibility of data science projects is made possible by a knowledge repository that contains the resources needed for the project at different levels of abstraction. In addition, features that support the use of the knowledge repository, such as a search function and version control, must be implemented in the knowledge repository. Also, the actual use of the knowledge repository must be included in a data science process model.

Keywords: Knowledge management, Data science projects, Knowledge repository, Reproducibility

ALKUSANAT

Diplomityön tekeminen oli palkitsevaa, mutta myös ajoittain raskasta. Olen kiitollinen tutkimuksen kohdeyritykselle kiinnostavan aiheen innoittamisesta ja arvokkaasta tuesta työn tekemisen aikana. Iso kiitos erityisesti myös haastateltaville mielenkiintoisista keskusteluista. Koen, että pääsin työn aihepiirien kautta kokoamaan keskeisimpiä oppejani tietojohtamisen opintojen eri vaiheista ja pohtimaan niiden merkitystä käytännössä.

Kiitos ohjaajalleni Henri Pirkkalaiselle kaikesta saamastani tuesta. Henriin oli aina helppo olla yhteydessä ja hänen kommenttinsa ohjasivat työn hyvin alkuun ja tarvittaessa oikeaan suuntaan. Lisäksi haluan kiittää perhettäni, läheisiäni ja ennen kaikkea avopuolisoani Juliaa koko yliopisto-opintojen aikana saamastani tuesta.

Kuopiossa, 5.5.2023

Juuso Halonen

SISÄLLYSLUETTELO

KUVALUETTELO

LYHENTEET JA MERKINNÄT

CRISP-DM	Yleinen datatieteellisen projektin prosessimalli (engl. Cross Industry Standard Process for Data Mining)
SECI-malli	Tiedonkonversiomalli, jonka vaiheiden kautta hiljainen tieto muuttuu eksplisiittiseksi ja takaisin hiljaiseksi (engl. Socialization, Externalization, Internalization, Combination)

1. JOHDANTO

Datatieteellisiä projekteja toteutetaan useilla toimialoilla ja niillä tavoitellaan tukea päätöksenteolle muun muassa raporttien ja muiden vastaavien datatuotteiden kautta (Cao, 2017). Datan määrän kasvu ja analysoinnin jatkuvasti kehittyvät kyvykkyydet ovat korostaneet datatieteellisten projektien haasteita esimerkiksi toistettavuuden osalta (Martinez et al., 2021). Heikko toistettavuus tulee huomioida, koska se alentaa luottamusta datatieteellisten projektien tuloksiin (Saltz, 2015). Toistettavuuden huono taso vaikeuttaa myös tiedon säilyttämistä ja oppimista aiemmista projekteista (Martinez et al., 2021). Sen lisäksi, että toistettavuus nähdään datatieteessä haasteena, tarjoaa se käytännössä myös avaimet laajempaan liiketoimintaan. Toistettavuus on mielenkiintoinen asia etenkin yrityksille, joiden ydinosaamista on datatieteellisten projektien toteuttaminen useissa eri liiketoimintaympäristöissä, sillä parempi toistettavuus tuo kyvyn palvella suurempaa määrää toimijoita, mikä puolestaan tekee yrityksen liiketoiminnasta tuottoisampaa.

Datatieteellisten projektien toistettavuusongelman ratkomiseksi Martinez et al. (2021) kääntyvät tietämyksenhallinnan puoleen ja esittävät alustavan ehdotuksen tietopankkien hyödyntämisestä tässä. Yhtä lailla analytiikkaprojektien parantamiseksi on ehdotettu tietopankin käyttöä tulevia projekteja ajatellen (Gökalp et al., 2022). Näissä ehdotuksissa on perää, sillä tietopankkien on tutkittu toimivan toistoa ja resurssien uudelleenkäyttöä vaativissa tilanteissa (Subramani et al., 2021) myös projekteja toteuttavissa organisaatioissa (Lindner & Wald, 2011). Lisäksi uudelleenkäytettävien resurssien määrittely on tunnistettu onnistuneiden datatieteellisten projektien taustatekijäksi (Berinato, 2019), mikä edelleen puhuu tietopankin hyödyntämisen puolesta.

Koska Martinezin et al. (2021) ehdottama tietopankki ja Gökalpin et al. (2022) suosittelema vastaava järjestelmä jäävät vain maininnan tasolle, ja tietopankin hyödyntämiselle on perusteet, tarkempi kuvailu sellaisesta on tarpeen. Tämä tutkimus pyrkii selvittämään tätä tarkempaa kuvailua ja sen tueksi muodostetaan päätutkimuskysymys ”*millainen tietopankki mahdollistaa datatieteellisten projektien toistettavuuden?”.* Alatutkimuskysymykset, joiden avulla pyritään vastaamaan päätutkimuskysymykseen, esitellään seuraavaksi.

Tietopankin käyttämisessä varsinainen sisältö on kaiken keskiössä ja sen laatuun tulee panostaa (Subramani et al., 2021; Veeravalli & Vijayalakshmi, 2021). Näin ollen ensimmäinen alatutkimuskysymys kohdennetaan siihen: *”mitä tietopankin tulisi sisältää, jotta se palvelee datatieteellisissä projekteissa toimimista?”.*

Tietämyksenhallinnan mahdollistava tietopankki on tarkkaan käyttötarkoitukseen suunniteltu tietojärjestelmä. Täten tietojärjestelmän tekniset ominaisuudet ovat relevantteja, joten ne huomioidaan toisessa alatutkimuskysymyksessä: *”mitä ominaisuuksia tietopankista tulee löytyä?”*.

Davenportin (2015) havainnoima historiallisten tietopankkien alhainen käyttöaste on huolestuttava. Tämän vuoksi tässä tutkimuksessa pyritään viimeisen alatutkimuskysymyksen avulla huomioimaan käyttöön vaikuttavia tekijöitä: *”mitkä tekijät ajavat tietopankin käyttöä?”.*

Tutkimus toteutetaan kartoittavana sekä teoriaa täydentävänä tapaustutkimuksena data-analytiikan ja datatieteellisten projektien parissa toimivan yrityksen kontekstissa. Aineisto kerätään haastattelemalla yrityksen työntekijöitä, jotka tutkimuksen näkökulmasta ovat tietopankin mahdollisia käyttäjiä. Kerätty aineisto analysoidaan teemoittelun avulla, missä tulokset jaotellaan teemoihin, jotka ohjaavat tutkimuskysymyksiin vastaamista.

Tutkimuksessa tavoitellaan kuvausta tietopankista, joka mahdollistaa datatieteellisten projektien toistettavuuden. Alatutkimuskysymysten mukaan kuvauksen tulee käsitellä ennen kaikkea tietopankin sisältöä, teknisiä ominaisuuksia ja käyttöön vaikuttavia tekijöitä. Viimeisen alatutkimuskysymyksen perusteella tutkimukselta odotetaan lisäksi tuloksia, jotka vahvistavat Ribièren ja Calabresen (2016) huomiot siitä, että teknologiat ja järjestelmät eivät yksinään ole ratkaisu tietämyksenhallinnan onnistumiselle.

Diplomityö lähtee liikkeelle taustateorian esittelyllä. Teoria on jaettu selkeästi kahteen kokonaisuuteen: tietämyksenhallinta ja datatieteelliset projektit. Tietämyksenhallinnan luvussa käsitellään alkuun yleisellä tasolla tietämyksenhallintaa ja taustalla vaikuttavia strategioita, joista kodifiointistrategia on tässä tutkimuksessa läsnä. Lisäksi esitellään keskeisiä tietämyksenhallinnan prosesseja ja niiden linkittymistä liiketoimintaan. Tästä edetään näitä prosesseja tukevien tietopankkien käsittelyyn ja selvennetään tietopankkien onnistumisen edellytykset.

Datatieteellisten projektien osalta tarkastelu alkaa projektien etenemisestä, joka kuvataan suositulla CRISP-DM prosessimallilla. Kuvailun jälkeen käsitellään projektin mahdollistavaa infrastruktuuria ja lopuksi paneudutaan tarkemmin datatieteellisten projektien toistettavuusongelmaan. Teorian jälkeen työ etenee tutkimusmenetelmän kuvailuun. Se on jaoteltu metodologisiin valintoihin, kohdeyrityksen esittelyyn ja tarkempaan kuvailuun

tutkimuksen toteutuksesta, aineiston keräämisestä ja analysoinnista. Kattavien tulosten aikaansaamiseksi keräämisen ja analysoinnin tukena hyödynnetään tietojärjestelmien onnistumiseen perustuvaa teoreettista viitekehystä.

Tuloksia tarkastellaan tutkimusmenetelmän kuvailun jälkeen. Tässä luvussa tulokset ovat jaoteltuna relevantteihin onnistuneen tietojärjestelmän kokonaisuuksiin, joista luodaan vielä yhteenveto luvun lopussa. Diplomityön viimeinen luku on pohdinta. Pohdinnassa tarkastellaan työn tuloksia teoreettisten sekä käytännön kontribuutioiden osilta. Lisäksi tässä luvussa arvioidaan tutkimusta kokonaisuudessaan, esitetään arvio tutkimuskysymyksiin vastaamisesta ja pohditaan mahdollisia jatkotutkimustarpeita.

2. TIETÄMYKSENHALLINTA

Tietoa ja tietämystä syntyy, kun saatavilla olevaa informaatiota arvioidaan sekä peilataan omien kokemusten kautta ja lopulta hyödynnetään päätöksenteossa. (Davenport & Prusak, 1998; Kidwell et al., 2000) Organisaatioiden sisällä tällaisia päätöksentekotilanteita tapahtuu päivittäin pienten valintojen muodossa. Se miten aiemmin kerrytettyä tietoa saadaan hyödynnettyä organisaatioissa siten, ettei jokaisen työntekijän tarvitse käydä samaa, mahdollisesti työlästä, uuden tiedon luomisen prosessia, on haastavaa (Szulanski, 1996). Etenkin, kun tietoa kertyy organisaatioissa useiden ihmisten toimesta jatkuvasti.

Systemaattista tapaa, jolla pyritään hallitsemaan organisaation ja sen työntekijöiden tietämystä kutsutaan tietämyksenhallinnaksi (Bergeron, 2003). Tietämyksenhallinta tarkastelee kuinka organisaatiot hankkivat ja hyödyntävät tietoa toimintansa tehostamiseksi (Choo, 2016). Tietämyksenhallintaa toteutetaan suoraan esimerkiksi teknologioiden ja tietojärjestelmien kautta tai epäsuorasti mukauttamalla esimerkiksi organisaatiorakenteita ja johtamismenetelmiä tiedonjakamisen kulttuurille sopivimmiksi. (Hislop, 2013) Tietämyksenhallinnan toteuttamiselle on esitetty kaksi isoa strategista suuntausta Hansenin et al. (1999) toimesta. Vaikka julkaisusta on aikaa, tämä määrittely nähdään yhtenä tietämyksenhallinnan perusperiaatteista (A.F. Ragab & Arisha, 2013; Hislop, 2013).

Toisena näistä strategioista on kodifiointistrategia, jossa tietojärjestelmien hyödyntäminen tiedon uudelleenkäytössä on keskeisessä asemassa. Tässä näkökulmassa arvonluonnin katsotaan tapahtuvan, kun saatavilla olevaa eksplisiittistä tietoa käytetään uudelleen. Kodifiointistrategiaa seuraava organisaatio saa yhtä lailla kilpailuetua eksplikoidun tiedon uudelleenkäytöstä. (Hansen et al., 1999) Tämän työn kontribuutio painottuu enemmän kodifiointistrategian puoleen, sillä tarkastelu ja pohdinta sopivasta tietopankista asettaa tiedon uudelleenkäytön isoon rooliin.

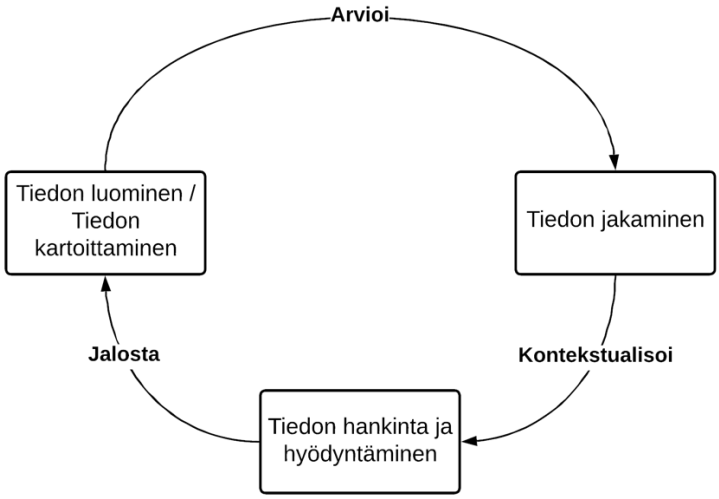
Kodifioinnin rinnalla toinen strategia on personointi. Tässä keskiössä on tiedon jakaminen ihmisten välisessä vuorovaikutuksessa. Tietojärjestelmät nähdään personointistrategiassa lähinnä välineenä tiedonjaon tehostamisessa esimerkiksi uusien viestintäpalveluiden muodossa. Toisin kuin kodifiointistrategiassa, personointistrategiaa noudattavan organisaation katsotaan saavan kilpailuetua uuden tiedon luomisesta, jota vauhdittaa henkilökohtaisen tiedon jakaminen. (Hansen et al., 1999)

Tietämyksenhallinnalla tavoitellaan muun muassa tehokasta tiedon jakamista ja tiedon uudelleenkäyttöä organisaation sisällä. Jotta tämä onnistuu ja on toimiva käytännössä,

tietämyksenhallinta pyrkii seuraamaan tiettyjä prosesseja. Niin kutsuttuja tietämyksenhallinnan prosesseja ja niiden kytköstä liiketoimintaprosesseihin esitellään tarkemmin seuraavassa alaluvussa.

2.1 Tietämyksenhallinnan sykli ja prosessit

Tietämyksenhallinnalle on määritelty prosesseja, jotka yhdessä luovat tietämyksenhallinnan syklin, joka puolestaan voidaan nähdä henkilökohtaisen tiedon reittinä koko organisaation strategiseksi voimavaraksi (Dalkir, 2011). Syklejä on esitetty useita, joiden pohjalta Dalkir (2011) on tuottanut kokoavan ja mahdollisimman yksiselitteisen syklin.



Kuva 1. *Tietämyksenhallinnan sykli (Dalkir, 2011)*

Kuvan 1 mukaisessa syklissä edetään tiedon luomisesta sen jakamiseen ja lopulta hyödyntämiseen. Luotu tai kartoitettu tieto arvioidaan merkityksellisyytensä osalta ennen kuin sitä jaetaan organisaatiossa eteenpäin sopivassa muodossa. Jaettua tietoa on lopulta mahdollista hyödyntää useammankin henkilön toimesta. (Dalkir, 2011)

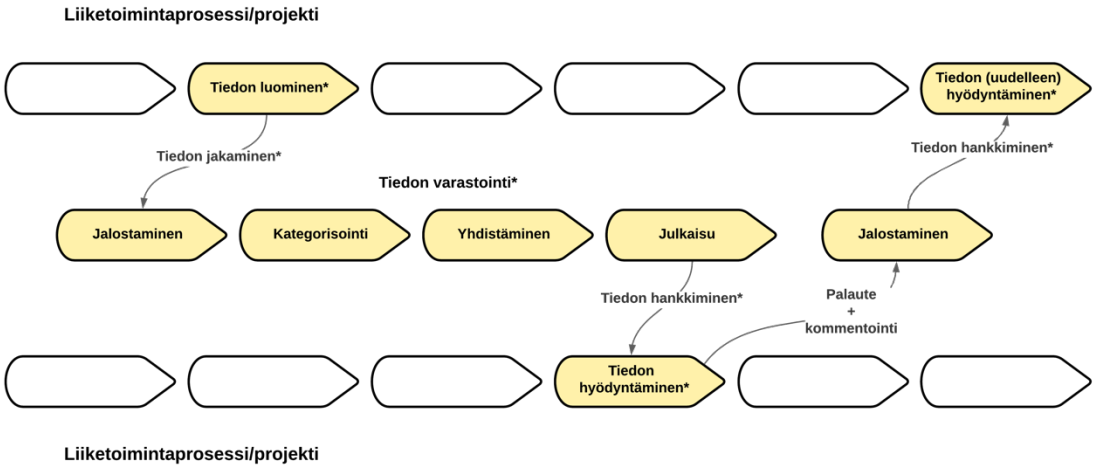
Kuten juuri havaittiin, tietämyksenhallinta elää käytännössä eri prosessien kautta. Keskeisiä tietämyksenhallinnan prosesseja, joita kuvasta 1 ja muusta kirjallisuudesta on noussut esiin ovat:

- **tiedon luominen** (Dalkir, 2011; Sahibzada et al., 2020; Raudeliuniene et al., 2021)
- **tiedon jakaminen** (Dalkir, 2011; A.F. Ragab & Arisha, 2013; Sahibzada et al., 2020; Raudeliuniene et al., 2021)
- **tiedon varastointi** (Sahibzada et al., 2020; Raudeliuniene et al., 2021)

- **tiedon hankkiminen** (Dalkir, 2011; A.F. Ragab & Arisha, 2013; Sahibzada et al., 2020; Raudeliuniene et al., 2021)
- **tiedon hyödyntäminen** (Dalkir, 2011; A.F. Ragab & Arisha, 2013; Sahibzada et al., 2020; Raudeliuniene et al., 2021)

Muitakin prosesseja kuten tiedon organisointi ja siirtäminen on aiemmin mainittu (Awad & Ghaziri, 2004), mutta ne voidaan perusperiaatteidensa perusteella katsoa kuuluvaksi edellä mainittuihin.

Edellisessä kuvassa syklin ja sen prosessien selkeää linkittymistä organisaation liiketoimintaan ei pääse seuraamaan. Tähän voidaan tuoda tueksi Maierin (2007) havainnollistama kuvaus siitä miten eri tietämyksenhallinnan prosessit ovat kytköksissä ja tukemassa tiedon uudelleenkäyttöä niin liiketoimintaympäristöjen tai projektien sisällä kuin useiden liiketoimintaympäristöjen ja projektien välillä. Alla olevassa kuvassa on esitetty kyseinen kokonaisuus, jossa Maierin (2007) mainitsemia vaiheita on uudelleennimetty tai tarkennettu vastaamaan edellä selvitettyjä keskeisiä tietämyksenhallinnan prosesseja.



Kuva 2. *Tietämyksenhallinnan prosessit kytkettynä liiketoimintaprosesseihin/projekteihin mukaillen Maier (2007)*

Kuvassa 2 on korostettu tietämyksenhallinnan prosesseja merkitsemällä niihin asteriski (*). Kuten kuvasta 2 nähdään, tietämyksenhallinta kokonaisuudessaan ja sen prosessit ovat kytköksissä useisiin liiketoimintaprosesseihin tai projekteihin. Tällä mahdollistetaan toisessa kontekstissa luodun tiedon hyödyntäminen myös muualla. Kuvan 2 mukainen tietämyksenhallinnan sitominen osaksi liiketoimintaprosesseja on tärkeää, jotta tiedon uudelleenkäytöstä saavutettu arvo saadaan realisoitua helpommin (Aviv et al., 2021).

Tiedon jakaminen, hankkiminen ja muut tietämyksenhallinnan prosessit eivät tapahdu ainoastaan ihmisten välisen vuorovaikutuksen kautta. Etenkin kodifioinnin näkökulmien kautta, jossa tieto voi olla olemassa itsenäisenä eksplikoituna olentona, prosessien tulisi tapahtua osana alustaa, joka on luotu tukemaan organisaation tietämyksenhallintaa ja toimintaa. Tästä näkökulmasta tällaisena alustana toimii jokin tietojärjestelmä tai kokoelma tietojärjestelmiä. (Hislop, 2013)

Tietämyksenhallinnan prosessien voidaan ajatella lähtevän liikkeelle uuden tiedon luomisesta. Uutta tietoa sitten jaetaan organisaatiolle paikkaan, jossa se varastoidaan. Varastoinnin yhteydessä tapahtuu muun muassa kategorisointia ja tiedon yhdistelyä (Maier, 2007; Hislop, 2013). Näiden jälkeen tieto on saatavilla ja sitä voidaan hankkia toisaalla. Hankittua tietoa hyödynnetään tiettyihin liiketoiminnallisiin tarpeisiin tarjoamaan vastauksia ja ratkaisuja. Hyödyntämisen jälkeen on tärkeää antaa palautetta käytetystä tiedosta, jotta sitä voidaan tarvittaessa jalostaa eteenpäin tai päivittää. (Dalkir, 2011)

Kodifiointistrategian taustalla vallitsee ajatus, että tieto voi olla eksplisiittisessä sekä hiljaisessa muodossa. Molemmille on aikansa ja paikkansa, ja molempia löytyy organisaatioista. Näitä muotoja ja liikkumista muotojen välillä, esitellään seuraavaksi.

2.2 Tiedon dualistinen luonne ja konversio

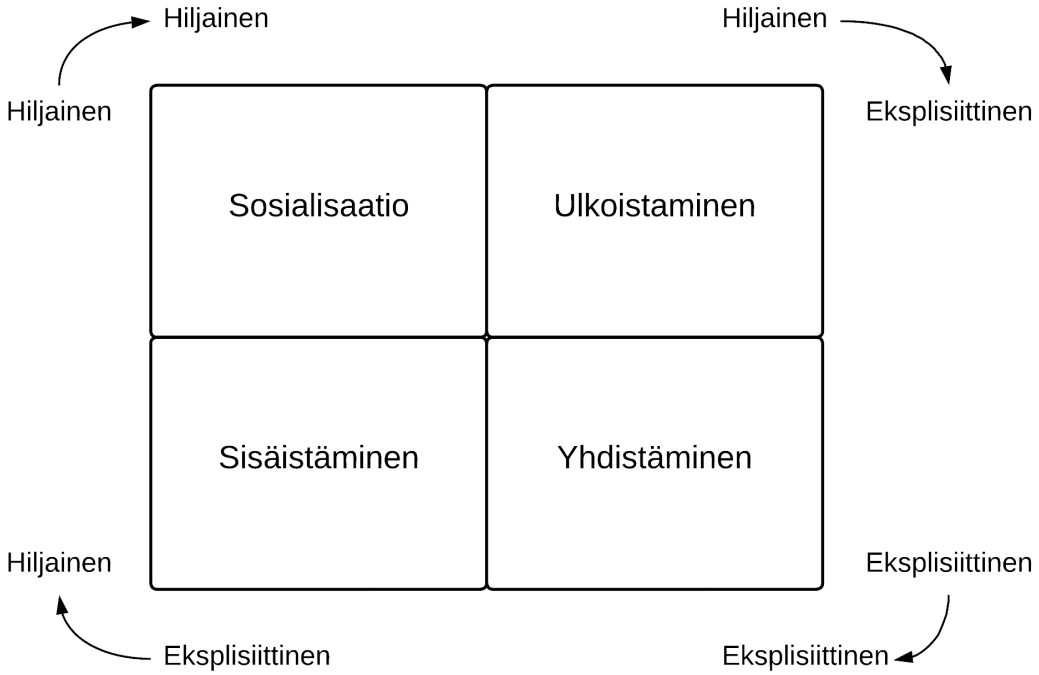
Tietämyksenhallinnassa tietoa luokitellaan usein dualistisesti joko hiljaiseksi tai eksplisiittiseksi (A.F. Ragab & Arisha, 2013; Hislop, 2013). Hiljaisena tietona pidetään tietämystä, jota on kertynyt henkilölle hänen kokemustensa kautta. Osa siitä on tiedostettua ja osa tiedostamatonta, ja se ohjaa vahvasti henkilön tapoja toimia. Eksplisiittistä tietoa on puolestaan tietämys, joka voidaan nähdä omana entiteettinään ja sitoutumattomana kehenkään henkilöön. (Polanyi, 1966) Eksplisiittistä tietoa voidaan esittää esimerkiksi kirjallisessa, matemaattisessa ja muussa konkreettisessa muodossa (Hislop, 2013).

Hiljaista tietoa on sen henkilökohtaisuuden ja osittaisen tiedostamattomuuden vuoksi hankala viestiä ja jakaa eteenpäin. Eksplisiittinen tieto on puolestaan kontekstista irrotettavaa ja helpommin jaettavaa. (Nonaka & Takeuchi, 1995; Hislop, 2013). Organisaatioissa on havaittavissa molempia tiedon lajeja ja seuraavaksi esitellään malli, joka ottaa nämä molemmat lajit huomioon pyrkimyksissään luoda ja jakaa uutta tietoa.

Kun uutta tietoa on luotu tai kokemuksia kerrytetty kuvasta 2 löytyvissä liiketoimintaprosesseissa tai projekteissa, sen luonne muuttuu, kun se etenee tietämyksenhallinnan syklin ja prosessien läpi. Se siirtyy hiljaisesta eksplisiittiseen ja aikanaan eksplisiittisestä

takaisin hiljaiseksi. Tätä muuntumista kuvaa hyvin Nonakan ja Takeuchin (1995) kehittämä tiedonkonversion malli.

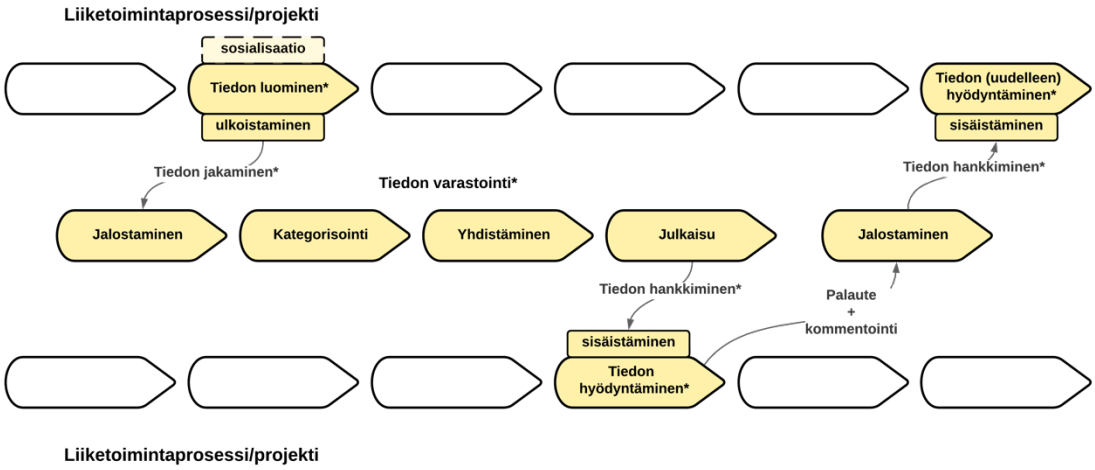
Vaikka Nonakan ja Takeuchin (1995) mallia on kritisoitu joiltain osin (Hislop, 2013), se on edelleen vahvasti mukana uusissa tutkimuksissa tiedon jakamisen ja luomisen ääreltä (Baldé et al., 2018; Goswami & Agrawal, 2022). Malli on edelleen ajankohtainen ja se käsittelee tiedon dualistista luonnetta selkeästi, minkä vuoksi juuri kyseinen viitekehys on tässä työssä mukana. Tiedonkonversiomalli, eli SECI-malli, koostuu neljästä eri vaiheesta. Kokonaisuudessaan malli on havainnollistettu seuraavassa kuvassa.



Kuva 3. *SECI-malli (Nonaka & Takeuchi, 1995)*

SECI-mallin vaiheet (kuva 3) ovat: sosialisaatio, ulkoistaminen, yhdistäminen ja sisäistäminen. Sosialisaatiossa ideana on hiljaisen tiedon välittyminen yksilöltä toiselle esimerkiksi keskustelun, havainnoinnin ja yhteistyön kautta. Ulkoistamisessa on kyse hiljaisen tiedon muuntamisesta eksplisiittiseksi. Sen tulisi tapahtua yksilötasolta ryhmätasolle muodoissa, jotka ovat käyttökelpoisia muille, kuten käsitteinä, malleina tai konsepteina. Seuraavana vaiheena on yhdistäminen, jossa tavoitteena on yhdistellä ulkoistamisen kautta syntynyttä eksplikoitua tietoa aiempaan eksplisiittiseen tietoon. (Nonaka & Takeuchi, 1995) Tämän vaiheen kautta on mahdollista syntyä muun muassa organisaatiotason parhaita käytänteitä ja standardeja (Hislop, 2013). Viimeisenä vaiheena SECI-mallissa on sisäistäminen. Sisäistäminen tapahtuu, kun yksilö toimii eksplikoidun tiedon

perusteella. Yksilö saa sisäistämisen kautta uusia kokemuksia ja mielikuvia, jotka sitoutuvat yksilöön hiljaisena tietona. (Nonaka & Takeuchi, 1995) SECI-mallin vaiheet voidaan kytkeä kuvan 2 esitettyihin prosesseihin, mikä havainnollistetaan seuraavassa kuvassa.



Kuva 4. *Tietämyksenhallinnan prosessit kytkettynä liiketoimintaprosesseihin/projekteihin SECI-mallin kanssa*

Kun uutta tietoa luodaan jossakin kuvan 4 liiketoimintaprosessissa tai projektissa esimerkiksi sosialisaation vauhdittamana, se täytyy muuttaa eksplisiittiseen muotoon, jotta se voidaan varastoida keskitettyyn säilytyspaikkaan (Hislop, 2013). On siis toteutettava ulkoistaminen. Varastoitu tieto prosessoidaan ja kategorisoidaan, eli SECI-mallin mukaan tapahtuu eksplisiittisen tiedon yhdistäminen. Kun varastoitua tietoa halutaan hyödyntää toisessa liiketoimintaprosessissa tai projektissa, eksplisiittinen tieto tulee sisäistää esimerkiksi käyttämällä eksplisiittistä tietoa työtehtävissä.

2.3 Tietopankit

Tietojärjestelmien rooli organisaation tietämyksenhallinnassa etenkin kodifiointistrategian perusteella on keskeinen (Hansen et al., 1999; Maier, 2007). Asianmukaisesti määriteltynä, suunniteltuna ja käyttöönotettuna, tietojärjestelmät ja tekniset alustat mahdollistavat onnistuessaan sujuvan tiedon jakamisen (Soto-Acosta & Cegarra-Navarro, 2016). Tietojärjestelmät toimivat muun muassa keskitettynä tiedon säilytyspaikkana ja kykenevät vastaamaan tallennetun tiedon kategorisoinnista. Lisäksi tietämyksenhallinnan tietojärjestelmät tarjoavat konkreettiset tavat ja väylät tiedon tallentamiseen sekä hakemiseen keskitetystä säilytyspaikasta. (Hislop, 2013)

Tietopankit (*engl. knowledge repository / knowledge base*) ovat olleet tunnistettuna yhdeksi isoksi kokonaisuudeksi puhuttaessa tietämyksenhallinnan toteuttamisen keinoista (Davenport & Prusak, 1998; Hislop, 2013). Tietopankit ovat tietojärjestelmiä, jotka sisältävät organisaatiolle keskeistä tietoa eksplisiittisessä muodossa ja täten toimivat keskitettynä tiedon säilytyspaikkana. Tietopankeilla tavoitellaan parempaa tiedon jakamista organisaation läpi, hiljaisen tiedon säilyttämistä työntekijöiden lähdettyä sekä liiketoiminnan tehostamista. Tietopankin tehtävänä on myös tarjota tietylle käyttäjäryhmälle ohjeita sekä resursseja standardisoiduista menetelmistä ja metodeista tiettyjen tehtävien suorittamiseen liittyen. (Sugumaran, 2016)

Seuraavassa luvussa käsitellään mikä tekee tietopankista onnistuneen ja sivutaan mitä haasteita tietopankkien ääreltä on historian saatossa ilmennyt. Onnistumisen tarkastelun yhteydessä esitellään DeLonen ja McLeanin (2003) kehittämä yleismaailmallinen tietojärjestelmien onnistumista arvioiva viitekehys, jonka rooli korostuu tämän työn tulosten analysoinnissa.

2.4 Onnistuneen tietopankin edellytykset

Yhdestä näkökulmasta voidaan ajatella, että mitä enemmän tietopankista löytyvää tietoa käytetään uudelleen, sitä onnistuneempi se on. Chhim et al. (2017) selvittivät mitkä tekijät vaikuttavat tiedon uudelleenkäyttöön elektronisesta tietopankista. Vahvin tilastollinen peruste heillä on löydölle, jonka mukaan käyttäjätyytyväisyyden ollessa korkea, tietoa käytettiin tietopankista enemmän uudelleen.

Käyttäjätyytyväisyyden nostamiseksi Chhim et al. (2017) kehottavat parantamaan tietopankin teknistä toteutusta muun muassa hakutoimintoa kohentamalla. Hakutoiminnon avulla pyritään selättämään pitkään vallinneet haasteet oikean tiedon löytämisen osalta, joita So ja Bolloju (2005) sekä Bock et al. (2010) ovat esitelleet. Chhim et al. (2017) myös suosittelevat hyödyntämään semanttisia teknologioita tiedon uudelleenkäytön aikaansaamiseksi. Semanttiset teknologiat pyrkivät löytämään merkityksiä datasta, dokumenteista sekä koodista tavoitellen sitä, että ihmisen lisäksi kone ymmärtää sisältöä (Sugumaran, 2016).

Tietopankin käyttäminen saattaa monissa tapauksissa olla uudenlainen tai ainakin erilainen lähestyminen työntekoon. Sen vuoksi on tärkeää, että johto on mukana mahdollistamassa tarvittavat käytännöt ja rutiinit se tukemiseen (Chhim et al., 2017), jotta tietopankin käyttö saadaan optimaaliseksi, eli osaksi normaalia työntekoa (Kankanhalli et al., 2011). Toisin kuin uusien lähestymistapojen ja muutoksien mahdollistamisessa, johdon

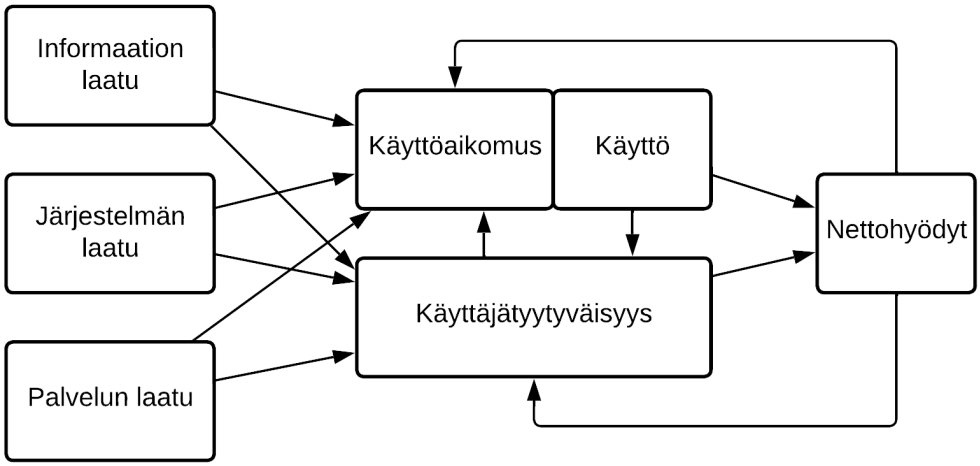
tuella ei varsinaisessa tiedon uudelleenkäytössä tietopankista ole merkittävää roolia (Veeravalli & Vijayalakshmi, 2021).

Ribière ja Calabrese (2016) painottavat, että dokumenttikeskeinen tietopankki epäonnistuu herkemmin, joten voidaan todeta, että onnistunut tietopankki pitää sisällään myös funktionaalista sisältöä. Oli sisältö dokumentteina tai funktionaalisessa muodossa, kuten toimivana koodina, tiedon laatu tietopankissa on yksi keskeinen menestystekijä. Laadun merkitys kasvaa huomattavasti, kun tallennettu tieto käsittää isoja kokonaisuuksia ja monimutkaistuu. Laadun lisäksi tallennetun tiedon pitää kattaa lähes kaikki siitä kyseisestä ilmiöstä, johon tallennettu tieto pureutuu. Muuten tiedon uudelleenhyödyntäjät joutuvat turvautumaan muihin resursseihin, mikä puolestaan kumoaa keskitetyn tietopankin tarkoituksen. (Filieri & Willison, 2016) Onnistumisen takaamiseksi Veeravalli ja Vijayalakshmi (2021) suosittelevat järjestelmän käyttöönoton alkuvaiheessa kohdistamaan palvelun vain niin sanotusti ”harvoille ja valituille”. Heidän mukaansa tämä helpottaa saamaan laajempaa käyttöä tietopankille tulevaisuudessa.

Kuten kaikkien tietojärjestelmien, niin myös tietopankkien kanssa voidaan onnistumista tarkastella DeLonen ja McLeanin (2003) kehittämän viitekehyksen avulla. Tässä luvussa ei kuitenkaan sisällytetä tietopankin onnistumisen tekijöitä malliin, vaan vasta esitellään se. Myöhemmin työn tulosten äärellä nämä kaksi kokonaisuutta linkittyvät syvemmin. DeLonen ja McLeanin (2003) julkaisema päivitetty versio tarkastelee onnistumista seuraavien dimensioiden kautta.

- Informaation laatu
- Järjestelmän laatu
- Palvelun laatu
- Käyttöaikomus & käyttö
- Käyttäjätyytyväisyys
- Nettohyödyt

Seuraavassa kuvassa on havainnollistettu miten edellä mainitut aspektit ovat kytköksissä toisiinsa.



Kuva 5. *Tietojärjestelmien onnistumiseen vaikuttavat tekijät (DeLone & McLean, 2003)*

Kuvan 5 purkaminen on selkeää tehdä vasemmalta oikealle. Informaation laatu kuvailee muun muassa tietojärjestelmän sisällön oikeellisuutta, ajankohtaisuutta ja kattavuutta. Järjestelmän laadun katsotaan puolestaan käsittelevän esimerkiksi järjestelmän helppokäyttöisyyttä, toiminnallisuutta ja joustavuutta. Järjestelmän laatu voidaan yhtä lailla ymmärtää tavoiteltujen ominaisuuksien kautta (Petter et al. 2013). Palvelun laadun avulla tarkastellaan tukitoimia järjestelmän ympäriltä ja sitä kuinka luotettavasti ja tarkasti käyttäjät saavat tukea. (DeLone & McLean, 2003)

Edellä esitetyt kolme dimensiota vaikuttavat kaikki käyttöaikomukseen sekä käyttäjätyytyväisyyteen. Käyttö merkitsee tässä mallissa järjestelmän käytön määrä, käyttötapaa tai tarkoituksenmukaisuutta. Ja jotta on käyttöä, siihen tarvitaan käyttöaikomus, joka syntyy edellisten dimensioiden onnistumisen kautta. Selkeyden vuoksi käyttöaikomus ja käyttö pidetään jatkossa yhdessä. Käyttäjätyytyväisyys merkitsee nimensä mukaisesti käyttäjien tyytyväisyyttä järjestelmää kohtaa. Käyttäjätyytyväisyyttä käsitellään mallissa esimerkiksi sen kautta kuinka usein käyttäjät palaavat käyttämään järjestelmää. Myös muut vapaamuotoisemmat mittarit auttavat tarkastelussa. (DeLone & McLean, 2003)

Yhdessä käyttäjätyytyväisyys ja käyttö -dimensiot vaikuttavat järjestelmästä saataviin nettohyötyihin joko positiivisesti tai negatiivisesti. Hyötyjä voivat olla muun muassa ajalliset säästöt, rahalliset säästöt, parantunut päätöksenteko ja eteneminen kohti strategisia tavoitteita. Kuten kuvan 5 mallista voidaan havainnoida, koetut hyödyt vaikuttavat takaisin käyttäjätyytyväisyyteen ja käytön aikomukseen. Positiivisena koetut hyödyt mahdollistavat järjestelmän käytön jatkuvuuden ja negatiivisena koetut hyödyt puolestaan laskevat järjestelmän käyttöastetta. (DeLone & McLean, 2003)

Tietopankin onnistumiseen spesifejä tekijöitä ei DeLone ja McLeanin (2003) viitekehyksestä suoranaisesti ilmene, mutta se on yleisesti toimivaksi todettu malli tarkastelemaan minkä tahansa tietojärjestelmän onnistumista. Kyseistä viitekehystä on myös aiemmin käytetty tietämyksenhallinnan ja tietopankkien kanssa esimerkiksi Filieri ja Willison (2016) toimesta. Heidän mallinsa oli johdannainen DeLonen ja McLeanin (2003) mallista ja otti vahvasti kantaa etenkin tietopankin sisällön laadullisiin tekijöihin ja niiden vaikutuksesta tiedon uudelleenkäyttöön. Wu ja Wang (2006) ovat tätä aiemmin myös sovittaneet DeLonen ja McLeanin (2003) mallia erityisesti tietämyksenhallinnan tietojärjestelmien tarkasteluun.

Syynä ettei tässä työssä sovelleta empiirisessä osuudessa kumpaakaan rajattua teoreettista viitekehystä on se, että tuoreessa kirjallisuudessa on painotettu uudenlaisten teknologioiden ja sovelluksien merkitystä tietämyksenhallinnalle (Nakash & Bouhnik, 2021), joten yleisemmän tason viitekehyksen katsotaan tuovan tuoreita – kenties tervetulleita – näkökulmia aiempien tietämyksenhallinnan sudenkuoppien välttämiseksi. Näistä nousee silti tärkeitä ja huomioitavia tietopankin onnistumisen tekijöitä: tietopankin tulee olla helposti integroitava, tietopankin tulee olla joustava, sisällön tulee olla laadukasta ja käytön hyötyjen tulee olla helposti havaittavissa (Wu & Wang, 2006; Filieri & Willison, 2016).

Vaikka tietojärjestelmät ovat yksi keskeinen kokonaisuus tietämyksenhallinnan toteuttamisen keinoista, tietämyksenhallinnan onnistuminen on riippuvainen muustakin kuin toteutetun tietojärjestelmän, kuten tietopankin onnistumisesta. Taustalla vaikuttavat teknologioiden lisäksi kulttuuri, hyötyjen mittaus, strategia, organisaatiorakenne, hallinto ja johtajuus sekä tietämyksenhallinnan standardien tai ymmärryksen taso. (Ribière & Calabrese, 2016) Lisäksi ihmisten välisellä keskustelulla on erittäin suuri vaikutus (Russell et al., 2016). Keskustelu johtaa yhteisöihin, jossa puidaan esimerkiksi ratkaisuja tiettyjen teknologioiden ääreltä, millä puolestaan saadaan aikaan uuden tiedon luomista osallistujien välillä SECI-mallin sosialisaatio vaiheen mukaisesti.

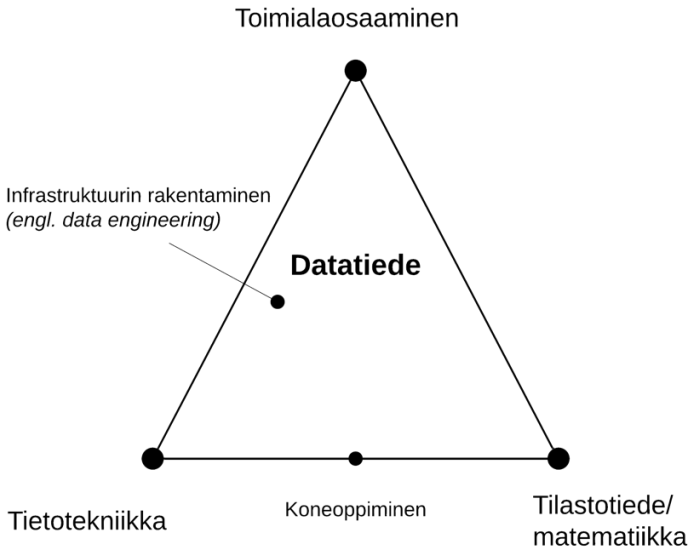
Yhteisöjä voi olla useita eri aihepiireistä ja yhdessä ne luovat niin sanotun tietämyksenhallinnan ympäristön. Aiempaan teoriaan peilaten, yhteisöllä voidaan esimerkiksi tarkoittaa kuvan 2 ja 4 projektien äärellä toimivia ihmisiä. Tietämyksenhallinnan ympäristössä syntyy uutta tietoa luonnollisesti ja jatkuvasti, sillä osallistujat ovat omasta aidosta kiinnostuksestaan yhteisöissä mukana. (Russell et al., 2016) Sosialisaation eli hiljaisen tiedon välittymisen lisäksi tietämyksenhallinnan ympäristössä tulisi tapahtua SECI-mallin muita vaiheita, jotta uutta tietoa päästäisiin luomaan mahdollisimman paljon. Näiden vaiheiden tukena jokin tietämyksenhallinnan tietojärjestelmä on otollinen – etenkin, jos käsitellään eksplisiittistä tietoa.

Miten tietopankki toimii osana tietämyksenhallinnan ympäristöä? Kaikki osa-alueet linkittyvät toisiinsa tietämyksenhallinnan onnistumista tarkastellessa. Yhteisöissä käydyn keskustelun tulokset eksplikoidaan, jotta uusi tieto ei pysyisi ainoastaan hiljaisena ja sidottuina yksilöihin. Eksplikoidut tulokset tallennetaan tietopankkiin malleina, käsitteinä ja menetelminä, ja tätä tallennettua tietoa yhdistellään aiempaan tietoon. Keskitetty varastointi mahdollistaa muun muassa yhdessä yhteisössä syntyvän tietämyksen hyödyntämisen toisessa yhteisössä, mikä puolestaan luo jälleen otollisen alkutilanteen uuden tiedon luomiselle. Tietopankki on kuvattu osana tietämyksenhallinnan ympäristöä kuvien 2 ja 4 keskimmäisenä prosessina.

3. DATATIETEELLISET PROJEKTIT

Datatiede on monia muita tieteenalojen läpileikkaava ala. Se hyödyntää muun muassa tilastotieteen, tietotekniikan, viestinnän, johtamisen ja sosiologian oppeja pyrkimyksissään tarjota tukea päätöksenteolle luomalla oivalluksia suuresta määrästä aineistoa (Cao, 2017; Kelleher et al., 2021). Tuki päätöksenteolle konkretisoituu niin kutsuttujen datatuotteiden avulla. Datatuotteita voi olla monenlaisia, kuten ennusteita, suosituksia, malleja, järjestelmiä tai kokonaisia palveluita. Kaikille näille on kuitenkin yhteistä se, että ne ovat joko suunniteltu viestimään datasta jalostettu informaatio ja tieto, tai ne ovat suunniteltu datasta jalostuneen informaation tai tiedon perusteella. (Cao, 2017) Oli datatuote minkälainen tahansa, tavoite päätöksenteon tuelle tai jopa automatisoidulle päätöksenteolle on aina taustalla.

Datatieteessä on keskeistä myös aina kyseisen toimialan ydinosaaminen, jotta datan avulla kyetään vastaamaan oikeisiin ja liiketoimintaa kiinnostaviin kysymyksiin (Davenport & Patil, 2012; Dhar, 2013; Berinato, 2019). Yhdessä toimialaosaamisen kanssa, aiemmin mainitut tieteenalat luovat datatieteen kokonaisuuden. Datatieteen osa-alueet ovat selvyyden vuoksi koostettu seuraavaan kuvaan.



Kuva 6. *Datatieteen osa-alueita mukaillen Emmert-Streib et al. (2016)*

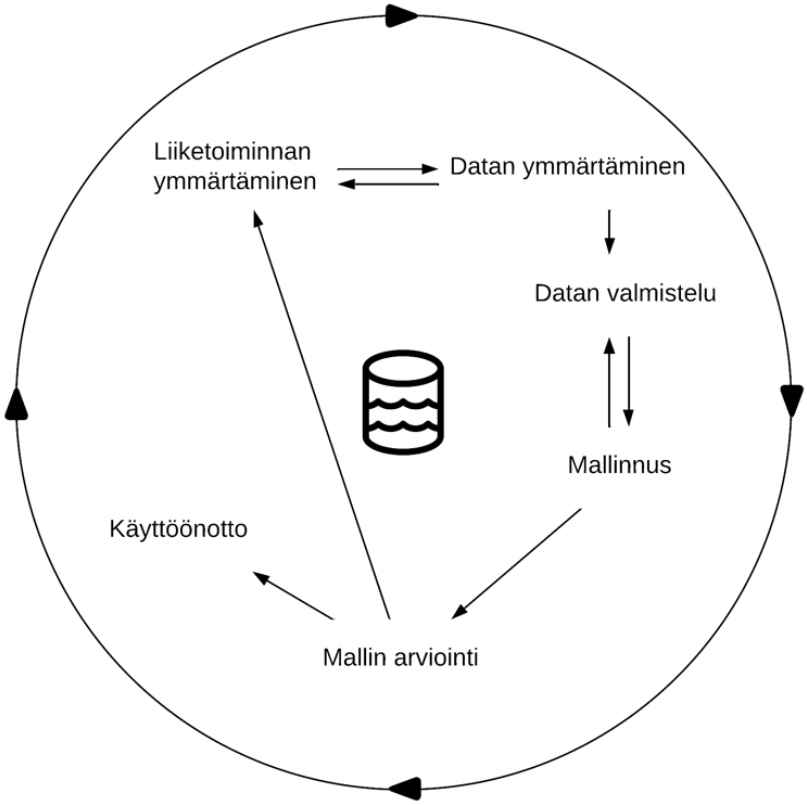
Kuvasta 6 nähdään, että kokonaisuudet kuten koneoppiminen ja infrastruktuurin rakentaminen ovat myös osa datatiedettä. Näiden lisäksi terminologisesti lähellä oleva data-analytiikka voidaan lukea osaksi tätä joukkoa (Cao, 2017). Siinä missä data-analytiikka

keskittyy merkityksien etsimiseen datasta (Cao, 2017), datatiede huomioi muun muassa kokonaisvaltaisia prosesseja, datalähteiden ominaisuuksia ja analyyttisten tulosten käyttöönottoa laajemmalla skaalalla (Provost & Fawcett, 2013). Tämän työn laajuudessa datatiede käsite pitää sisällään kuvan 6 mukaisia kokonaisuuksia ja kaikkea niiden väliltä.

Käytännössä datatiedettä harjoitetaan datatieteellisten projektien kautta, joita toteutetaan toimialoista riippumatta. Usein datatieteelliset projektit seuraavat ennalta määriteltyjä prosessimalleja sekä koostuvat samantyyppisistä projekteja tukevista infrastruktuuriratkaisuista. (Kelleher et al., 2021) Projektien eteneminen prosessimallien kautta ja taustalla toimiva infrastruktuuri ovat seuraavien alalukujen käsiteltävät teemat. Lisäksi viimeisessä alaluvussa tarkastellaan toistettavuuden haasteellisuutta datatieteellisten projektien ääreltä.

3.1 Projektin eteneminen

Datatiede pyrkii liikkumaan aineistosta kohti päätöksentekoa tukevia oivalluksia. Tähän etenemiseen on kehitetty paljon erilaisia prosessimalleja, mutta laajalti käytössä on edelleen CRISP-DM (*Cross Industry Standard Process for Data Mining*), jonka suosio perustuu ainakin osittain sen riippumattomuuteen toimialoista ja teknologioista. (Kelleher et al., 2021) CRISP-DM prosessi koostuu kuudesta eri vaiheesta, joita ovat liiketoiminnan ymmärtäminen, datan ymmärtäminen, datan valmistelu, mallinnus, mallin arviointi ja käyttöönotto. (Chapman et al. 2000)



Kuva 7. *CRISP-DM prosessimalli (Chapman et al., 2000)*

CRISP-DM prosessi alkaa liiketoiminnan ymmärtämisestä ja etenee kuvan 7 nuolten mukaisesti. Suoritusjärjestys voi kuitenkin tapauskohtaisesti muuttua ja vaiheiden välillä voidaan liikkua edestakaisin. Liiketoiminnan ymmärtäminen ja datan ymmärtäminen -vaiheissa määritellään tavoitteet, joihin data-analytiikan avulla pyritään vastaamaan. Dataa käsittelevän henkilön tehtävänä on näissä vaiheissa ymmärtää liiketoiminnan tarpeet ja pohtia onko käsillä oleva aineisto soveltuva liiketoiminnan tarpeisiin tarvittavan ratkaisun kehittämiseksi. (Chapman et al. 2000; Kelleher et al., 2021)

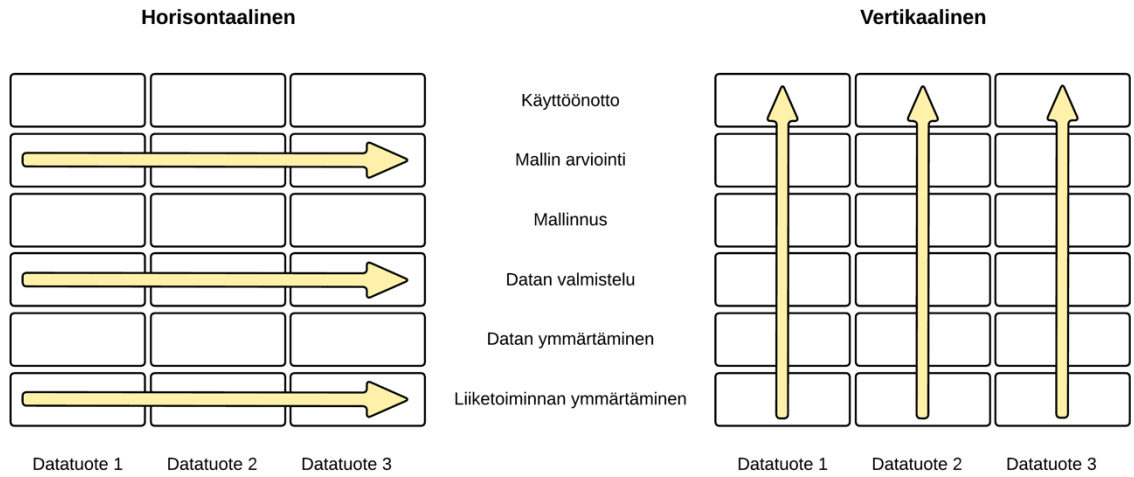
Datan valmistelu -vaiheen avulla pyritään luomaan analyysin mahdollistava aineisto (Kelleher et al., 2021). Vaihe pitää sisällään datan siivoamista, rakentamista, integroimista ja formatointia. (Chapman et al. 2000) Siivoamisella voidaan tarkoittaa esimerkiksi duplikaattien poistamista, irrelevanttien sarakkeiden poistoa tai tiettyjen arvojen muuntamista tai korvaamista. Arvoja voidaan esimerkiksi muuttaa merkkijonosta päivämäärämuotoiseksi.

Mallinnusvaihe pitää sisällään hyödynnettävien mallinnusmenetelmien valitsemisen sekä mallien rakentamisen. Kirjava joukko koneoppimisen menetelmiä ja algoritmeja, jotka eristävät aineistosta hyödyllisiä havaintoja, ovat tämän vaiheen keskiössä.

Mallinnus sisältää tyypillisesti useiden eri menetelmien kokeilua, minkä vuoksi tulee huomioida, että mallinnusvaihe on vahvasti sidoksissa datan valmisteluun, sillä eri menetelmät voivat vaatia dataa eri muodoissa. Mallinnus nostaa myös usein esiin virheitä datasta, joten valmisteluvaiheeseen on palattava. (Kelleher et al. 2021)

Arviointivaiheessa arvioidaan ja dokumentoidaan mallinnusvaiheen kautta saatuja tuloksia. Arvioinnissa on olennaista huomioida, etteivät tulokset välttämättä vastaa liiketoiminnan sille asettamia tavoitteita. Tässä vaiheessa datatieteellinen prosessi voidaan joko aloittaa täysin alusta tai tunnistaa vaihe, jonka johdosta tavoitteiseen ei olla päästy. Arvioinnin tuloksena voi myös olla päätelmä, jonka mukaan tulokset eivät ole olennaisia alkuperäisen liiketoiminnan tavoitteen osalta ja siten prosessin jatkaminen olisi turhaa. Viimeisessä vaiheessa malli otetaan käyttöön jossakin ympäristössä, jossa myös loppukäyttäjät voivat hyödyntää sitä. Tässä vaiheessa suunnitellaan miten malli yhdistetään organisaation muuhun tekniseen infrastruktuuriin ja liiketoiminnan prosesseihin. Lisäksi käyttöönoton yhteydessä tehdään päätöksiä mallin monitoroinnista ja ylläpidosta. (Chapman et al. 2000; Kelleher et al., 2021)

Liiketoimintaympäristössä CRISP-DM prosessimallia voidaan suorittaa sekä horisontaalisesti että vertikaalisesti lopputuotteina syntyviä datatuotteita ajatellen. Horisontaalisessa tavassa, joka vastaa enemmän vesiputousmallia ohjelmistokehityksen puolelta, CRISP-DM suoritetaan vaihe vaiheelta kaikille tavoitelluille datatuotteille. Vertikaalisessa tavassa puolestaan edetään datatuote kerrallaan kaikki vaiheet läpi, mikä taas vastaa enemmän ketterää mallia ohjelmistokehityksen puolelta. (Hotz, 2023) Nämä kaksi tapaa ovat havainnollistettuna alla.



Kuva 8. *CRISP-DM toteuttamistavat (Hotz, 2023)*

Suoraviivaisista etenemistä kuvaavista nuolista huolimatta, kuvassa 8 mallin toteuttaminen on kuitenkin iteratiivista taustalla olevan CRISP-DM prosessimallin mukaisesti. Hotz

(2023) suosii blogitekstissään näistä kahdesta tavasta vertikaalista. Hän perustelee suositustaan muun muassa sillä, että sitä kautta datatieteellisen projektin arvo koetaan aiemmin ja mallin arviointiin päästään nopeammin käsiksi. Arvokasta palautetta loppukäyttäjiltä saadaan myös aiemmin ja se voidaan ottaa sujuvammin huomioon joko palattaessa yhden datatuotteen iteraatiossa vaiheita takaisin tai seuraavan datatuotteen kehittämisen parissa.

Vaikka CRISP-DM on vakiinnuttanut asemansa yhtenä suosituimpana prosessimallina, muita varteenotettavia malleja on alkanut syntyä. Katalyyttinä tässä synnyssä on toiminut esimerkiksi aineiston määrän suuri kasvu (Saltz, 2015) ja uudenlaisten datan muotojen sekä datalähteiden, kuten sensoridatan esilletulo (Nagashima & Kato, 2019). Lisäksi, kun modernit menetelmät kuten lean startup ja muotoiluajattelu yleistyvät ohjelmisto- sekä tuotekehityksessä, datatieteellisten projektien etenemistä ja kehittämistä on myös arvioitu uudelleen. Tämän perusteella on ehdotettu CRISP-DM:n jatkoksi uusia malleja, jotka keskittyvät etenkin joustavuuteen ja jatkuvuuteen (Ahmed et al., 2018).

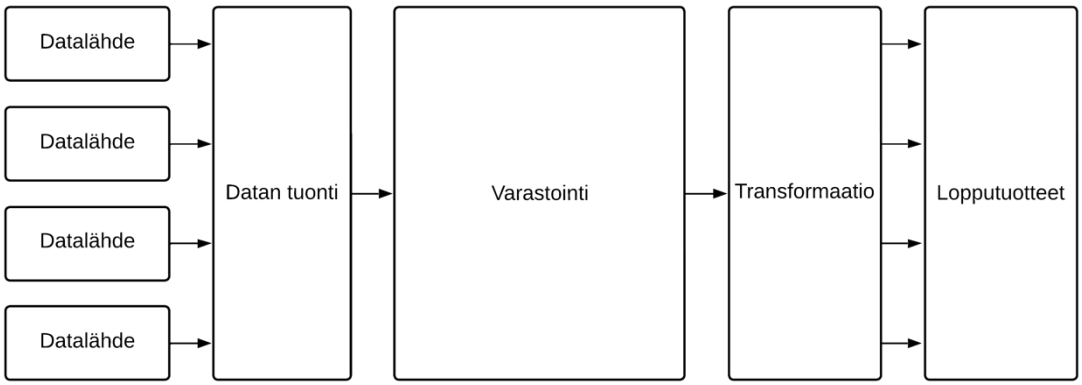
Koska CRISP-DM on kuitenkin yleisin malli datatieteellisen projektin toteutukseen ja uudet mallit ovat lähes poikkeuksetta johdannaisia siitä (Rotondo & Quilligan, 2020), katsotaan CRISP-DM:n olevan riittävä rajaus tämän työn datatieteellisien projektien ja taustalla vaikuttavia prosessimallien käsittelyyn. Erittäin olennainen huomio myös tämän työn tulosten kannalta on se, että organisaatiot käyttävät usein myös omia prosessimallejaan (Saltz, 2015). Näin organisaatiokohtaiset parhaat käytänteet pääsevät hioutumaan selkeästi määritellyn yleismaailmallisen prosessin jatkeeksi.

3.2 Infrastruktuuri

Datatieteellinen prosessi tarvitsee toimiakseen joukon teknologioita ja palveluita. Tämä taustalle rakennettu infrastruktuuri tukee edellisessä alaluvussa esitellyn datatieteellisen prosessin monia vaiheita, kuten aineiston valmistelua ja mallintamista. Oikein valittu ja toteutettu infrastruktuuri tukee prosessin iteratiivista luonnetta ja mahdollistaa esimerkiksi uusien datatuotteiden kehittämisen vanhojen rinnalle saman infrastruktuurikokonaisuuden alle.

Datatieteellisten projektien infrastruktuureissa on havaittavissa yhteneväisyyksiä käytettyjen komponenttien ja välineiden osalta. Yleisesti ottaen infrastruktuuriratkaisut voidaan jakaa seuraavasti: datalähteet, varastointi sekä lopputuotteet (Kelleher et al., 2021). Bornstein et al. (2020) laajentavat tätä jaottelut sisältämään myös aineiston tuonnin ja transformaation omina kokonaisuuksinaan. Tuonti sijoittuu datalähteiden ja varastoinnin

väliin. Transformaatio sijoittuu puolestaan varastoinnin ja lopputuotteiden väliin. Infrastruktuurin ympärillä vaikuttavat myös muun muassa käyttöoikeudet, hallintakäytänteen ja monitorointi (Bornstein et al., 2020). Infrastruktuuriratkaisujen jako havainnollistetaan seuraavassa kuvassa.



Kuva 9. *Tyypillinen datatieteellisen projektin infrastruktuuri mukaillen Bornstein et al. (2020) ja Kelleher et al. (2021)*

Kuvassa 9 esitetyn infrastruktuurin datalähteisiin kuuluu esimerkiksi organisaation erilaiset ohjelmistot ja sovellukset tai sensoridataa tuottavat mittalaitteet. Varastoinnin tehtävänä on puolestaan keskitetysti kerätä lähteistä tulevaa dataa (datan tuonti) päätöksenteon tueksi ja mahdollistaa koko organisaation laajuinen tiedon jakaminen ja analysointi. Varastoitu data analysoidaan ja lopulta siitä saatu informaatio tarjotaan oikeassa formaatissa (transformaatio) lopputuotteen kuten raportin kautta sitä tarvitseville. (Kelleher et al., 2021)

Infrastruktuuriratkaisujen käyttökohteet ja -tarkoitukset ovat havaittavissa CRISP-DM prosessimallin taustalla. Esimerkiksi datan valmistelu vaihetta koskettaa datan tuonti, varastointi ja transformaatio. Lisäksi kun datan valmistelussa tulee tilanteita vastaan, jossa useita datasettejä pitää yhdistellä, oikeanlainen varastointiratkaisu on keskeinen, sillä yhdestä keskitetystä tietovarastosta aineistojen yhdistely on huomattavasti suoraviivaisempaa kuin vaikkapa suoraan ulkoisista järjestelmistä (Kelleher et al., 2021).

3.3 Toistettavuus

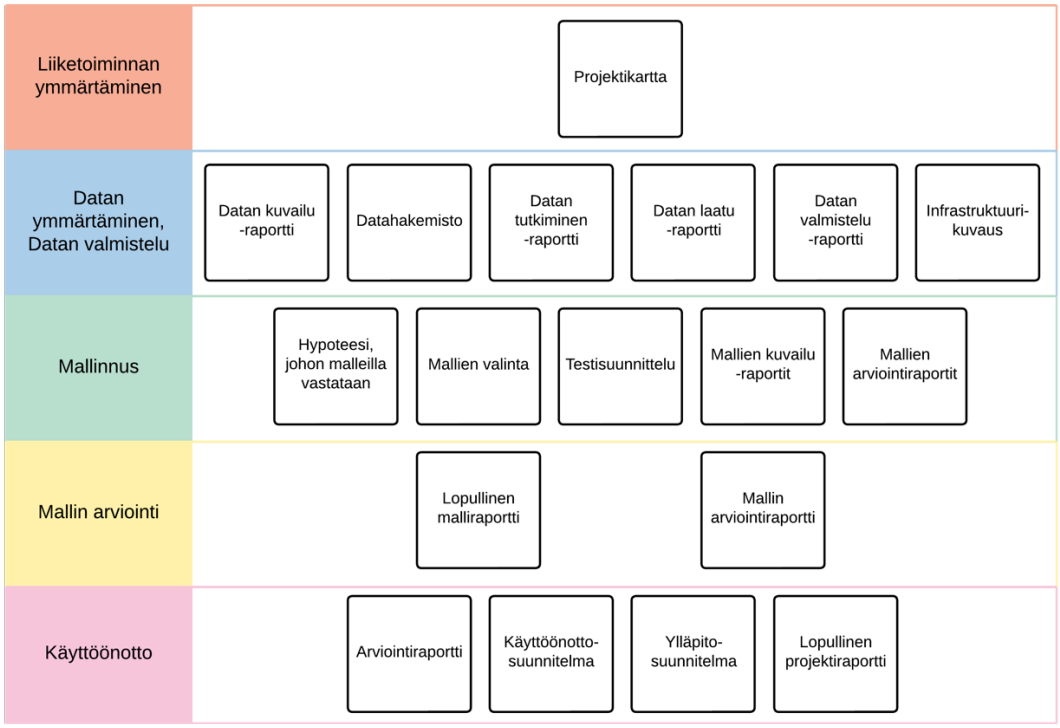
Datatieteelliset projektit ja niiden toteuttaminen ovat muuttumassa. Taustalla on tarve vastata monimutkaisempia liiketoiminnan vaatimuksia isommalla määrällä aineistoa ja tehokkaimmilla koneoppimisen algoritmeilla. Uudistuneet tarpeet tuovat datatieteellisille projekteille mukanaan haasteita, joita ilmenee tiimin, projektin sekä datan ja informaation ääreltä. Yhtenä esimerkkinä näistä haasteista on heikko toistettavuus (Martinez et al., 2021).

Toistettavuuden haasteellisuutta tarkastellaan työssä datatieteellisten projektien tasolla. Martinezin et al. (2021) mukaan sitä kuitenkin esiintyy myös paljon matalammalla tasolla, kuten analyysien tulosten kanssa. Kompleksisilla algoritmeilla ja menetelmillä on siis ajoittain vaikeuksia tuottaa samoja tuloksia kuin mitä on jo kertaalleen tuotettu, mikä alentaa luottamusta datatieteellisten projektien tuloksiin (Saltz, 2015).

Datatieteellisissä projekteissa yksittäisten tulosten ja mallien lisäksi tieto siitä miten kyseinen projekti eteni ja päätyi tiettyihin tuloksiin, on kiinnostavaa ja erittäin keskeistä, kun tarkastellaan toistettavuutta projektitasolla (Martinez et al., 2021). Pyritään siis selvittämään miksi mitäkin valintoja tehtiin projektin edetessä. Miksi suunniteltiin infrastruktuuri X? Miksi siinä paras tiedostoformaatti on Y? Valinnat ja niiden perustelut ovat vahvasti sitoutuneet hiljaisena tietoja niiden tekijöihin.

Tulevien projektien kannalta on kriittistä, että tieto saadaan jaettua laajemmin, sillä uuden datatieteellisen projektin parissa toimivat henkilöt eivät välttämättä ole samat henkilöt, ketkä päätyivät aiempiin valintoihin. Tarkoituksena on panostaa organisaatiotasolla tiedon säilymiseen ja oppimiseen, ja välttää tilanteet, jossa käytetään turhaa aikaa ratkomaan samoja ongelmia uudelleen. (Martinez et al., 2021) Kuten Martinez et al. (2021) sekä Gökalp et al. (2022) mainitsevat, keskitetyn tietopankin toteuttaminen pystyisi mahdollisesti vastaamaan tähän projektitason toistettavuuden haasteeseen. Tarkempi määrittely siitä millainen tietopankin tulisi olla selvitetään työn empiirisen osuuden avulla.

Haertel et al. (2022) tarttuvat myös Martinezin et al. (2021) toistettavuusongelmaan ja kokoavat kuvauksen datatieteellisien projektien artefakteista. Haertel et al. (2022) tuovat esiin kirjallisuuskatsauksen kautta CRISP-DM prosessin eri vaiheissa tuotettavat dokumentit, jotka ovat esitetty seuraavassa kuvassa laatikoiden sisällä.



Kuva 10. *Yleiskatsaus tuotettavista dokumenteista datatieteellisen projektin eri vaiheissa (Haertel et al., 2022)*

Tarkoituksena näillä kuvasta 10 löytyvillä dokumenteilla on tarjota mahdollisimman tarkka kuvaus tehdyistä asioista projektin aikana. Tarkkojen kuvausten avulla hyvin onnistunut datatieteellinen projekti on helpompi toistaa ja yhtä lailla epäonnistunut projekti välttää. Haertel et al. (2022) ovat loistavasti myös huomioineet vallitsevan nykytilan, jossa dokumentointi nähdään yrityksissä helposti taakkana, minkä vuoksi se jää usein alhaiselle tasolle. Tämän perusteella he kaipaavatkin jatkotutkimuksia joko artefaktien tuottamisen vaikutuksen viestimisestä tai tavoista automatisoida dokumentointia.

Tuotettava dokumentaatio datatieteellisessä projektissa nostettiin tähän teoriaosuuteen kahdesta syystä esiin. Ensimmäinen syy on kyseisen tutkimuksen olennaisuus tälle työlle, kun se pyrkii vastaamaan samaan ajankohtaiseen toistettavuusongelmaan. Toisena syynä on Haertelin et al. (2022) esittelemien raporttien ja muun dokumentaation mahdollinen linkittyminen tietopankin sisältöön, mitä analysoidaan tarkemmin luvussa pohdinta, kun tämän työn empiriaa verrataan aiempiin tutkimuksiin.

4. TUTKIMUSMENETELMÄ

Tässä kappaleessa esitellään metodologiset valinnat, jotka koottuna toimivat tutkimusmenetelmänä. Tarkoin määritelty tutkimusmenetelmä tarjoaa keinot tutkimuskysymyksiin vastaamiseen ja rakentaa muun muassa tutkimuksen luotettavuutta. Metodologisista valinnoista edetään kohdeyrityksen esittelyyn, minkä jälkeen paneudutaan aineiston keräämisen ja analysoinnin tarkempiin kuvailuihin.

4.1 Tutkimusmetodologia

Tutkimusmetodologiset valinnat ovat tarkemmin jaoteltuna tutkimusfilosofian, lähestymistavan, tutkimusstrategian, aineistonkeruun ja aineistonanalyysin valintoihin. Tähän työhön on valittu pragmaattinen suuntaus tutkimusfilosofian osalta. Pragmatismi on Elkjaerin ja Simpsonin (2011) mukaan käytännönläheinen ote, joka keskittyy ongelmanratkaisuun sekä kehittämään tulevia käytäntöjä, minkä vuoksi se sopii tämän tutkimuksen tarkoitukseen selvittää toimivan tietopankin edellytyksiä rajatussa kontekstissa.

Lähestymistavan valinnassa pitää puolestaan pohtia onko tavoitteena edetä aineistosta teoriaan (induktiivinen), testata teoreettisia käsitteitä aineiston avulla (deduktiivinen) vai luoda teoriaa kerätystä aineistosta ja myös testata sitä käytännössä (abduktiivinen) (Saunders et al., 2019). Tässä tutkimuksessa hyödynnetään induktiivista lähestymistapaa, sillä aineistoa käytetään työn tulosten lähtökohtana. Ja kuten sanottu, konteksti on rajattu, jonka vuoksi induktiivinen lähestymistapa on myös siltä osin soveltuva (Saunders et al., 2019). Lisäksi tämä aineistolähtöinen ote ohjaa tutkimuksia – kuten myös tätä työtä – laadullisen tutkimuksen piiriin (Eskola & Suoranta, 1998).

Tutkimusstrategiaksi esitetään tapaustutkimus, joka sopii, kun tarkastellaan jotakin ilmiötä tarkoin rajatussa ympäristössä, kuten tietyn yrityksen toimintaympäristössä (Yin, 2009). Tapaustutkimusta voi Vossin et al. (2002) mukaan käyttää etsintään, teorioiden rakentamiseen, teorioiden testaukseen ja teorioiden paranteluun. Esitetyt tapaustutkimuksen tarkoitukset sopivat myös hyvin induktiiviseen lähestymistapaan, mikä luo osaltaan vankkaa pohjaa yhtenäiselle ja selkeälle tutkimusmenetelmälle. Tässä työssä tapaustutkimus koskee lähinnä teorian parantelua. Tapaustutkimuksen valinta tähän työhön on otollinen, sillä kohdeyrityksessä on tunnistettu tarve kehittää tietämyksenhallintaa liiketoimintansa tueksi. Tapaustutkimuksen kautta voidaan tähän tarpeeseen tuoda lisää ymmärrystä tarkasteltavasta aiheesta ja ennen kaikkea tarjota arvokkaita havaintoja joh-

tamisen tueksi. Myös teoriaan voidaan kontribuoida tapaustutkimuksen avulla, kun ilmiötä päästään tutkimaan sen luonnollisessa ja todellisessa ympäristössä. Tällaisen tutkimuksen kautta saadaan arvokasta käytäntöön sidonnaista tietoa, jonka kautta merkityksellistä sekä relevanttia teoriaa voidaan luoda tai vahvistaa. (Meredith, 1998)

Edellä esitetyt valinnat johdattavat tutkimuksen lopulta aineistonkeruu- sekä analysointimenetelmien rajaukseen (Saunders et al., 2019). Tässä laadullisessa tutkimuksessa empiirinen osuus eli aineistonkeruu toteutetaan puolistrukturoituna teemahaastatteluna. Haastattelut ovat yleinen laadullisen tutkimuksen aineistonkeruumenetelmä ja hyvä tapa selvittää ajatuksia sekä mielipiteitä tietystä aiheesta pintaa syvemmältä (Eskola & Suoranta, 1998; Puusa et al., 2020) Tämän työn haastattelujen aihepiirit ovat teemahaastattelun tapaisesti ennalta määritettyjä ja lisäksi kysymykset ovat puolistrukturoidun haastattelun tapaan pääsääntöisesti avoimia. Vaikka Eskolan ja Suorannan (1998) mukaan valmiita kysymyksiä ei teemahaastattelussa varsinaisesti tarvita, jäsennellyn haastattelurungon toteuttaminen selkeyttää haastattelutapahtumaa ja myöhemmin aineiston analysointia. Teemahaastattelun on katsottu myös sopivan tapaustutkimukseen menetelmäksi (Saundersin et al., 2019).

Aineiston analysoinnin kolme selkeää vaihtoehtoa ovat aineistolähtöinen, teoriasidonnainen ja teorialähtöinen analyysi (Puusa et al., 2020). Tässä työssä aineiston analysointi tehdään induktiivisesta lähestymistavasta huolimatta teoriasidonnaisesti, jonka mukaan tulokset perustuvat aineistoon, mutta isommat kokonaisuudet, kuten teemat, ovat sidottavissa teoriaan (Tuomi & Sarajärvi, 2018).

4.2 Kohdeyritys

Kohdeyritys on tunnistanut selkeän tarpeen ja mahdollisuuden toteuttaa tehokkaampaa tiedon uudelleenkäyttöä datatieteellisissä projekteissa muun muassa tietopankin avulla. Kuten jo huomattu, sama tarve on tuoreessa kirjallisuudessa esillä, mutta määrittelyjä tarpeeseen toimivasta tietopankista ei ole kattavasti saatavilla. Tällaisten vähän tutkittujen ilmiöiden selvittämiseen tapaustutkimus on toimiva ja hyvä lähtökohta teorian laajentamiselle (Voss et al., 2002). Lisäksi tutkimuksen tuloksilla odotetaan olevan käytännön merkitystä kohdeyritykselle, joten kontekstin rajaaminen siihen tapaustutkimuksen kautta on Farquharin (2012) mukaan otollista.

Tapaustutkimuksen kohteena on asiantuntijaorganisaatio, joka palvelee asiakasyrityksiä useilla eri toimialoilla datatieteen ja sen osa-alueiden parissa. Kohdeyrityksen tavoitteena on tukea asiakkaidensa liiketoimintaa datan avulla. Kohdeyritys toteuttaa asiakkailleen sekä kokonaisia datatieteellisiä projekteja että yksittäisiä datatuotteita.

Kohdeyrityksen dataperusteinen tarjoama voidaan eritellä strategian, operatiivisen ja kulttuurin kokonaisuuksiin. Strategian osalta kohdeyritys pyrkii tunnistamaan kohteet, joihin data tuo lisäarvoa. Operatiivinen tarjoama pitää sisällään muun muassa viimeisimmillä teknologioilla toteutetut infrastruktuuriratkaisut. Kulttuuri puolestaan käsittelee keinoja, joiden avulla organisaatioita mukautetaan hyödyntämään datalähtöistä päätöksentekoa.

4.3 Tutkimuksen toteutus ja aineiston kerääminen

Teoriasidonnaisen analyysin sallimana haastattelurunko (liite A) rakennettiin ja kehitettiin sisältämään Delonen & McLeanin (2003) viitekehyksen mukaisia kokonaisuuksia onnistuneen tietojärjestelmän taustalla. Nämä kokonaisuudet päätyivät suurimmilta osin haastattelurungon otsikoiksi, mikä koettiin hyväksi jaotteluksi haastattelurungolle esiin nousevia teemoja ajatellen. Tässä jaottelun valinnassa tulee hyvin myös esiin työn pragmaattinen tutkimusfilosofia, jossa tutkijan uskomuksien katsotaan vaikuttavan tutkimukseen (Saundersin et al., 2019).

Tutkimuksen kulkua ohjaa päätutkimuskysymys, joka tässä tapauksessa pyrkii vastaamaan kysymykseen *millainen tietopankki*. Tähän vastaukseksi odotetaan laajaakin kuvausta, mutta eritoten teknisiä aspekteja, minkä vuoksi haastattelurungon jaottelu onnistuneen tietojärjestelmän kokonaisuuksiin on loogista. Haastattelurungossa tuotiin lisäksi esille tutkimuksessa rajattu konteksti datatieteellisille projekteille. Myös tietämyksenhallinnan prosessiluonnetta huomioitiin kysymyksissä etenkin tiedon jakamisen/tallentamisen ja tiedon hyödyntämisen osilta.

Otanta haastatteluihin tehtiin kohdeyrityksestä harkinnanvaraisesti. Tällaista otantaa käytetään usein laadullisessa tutkimuksessa, kun pyritään selvittämään asioita perusteellisesti (Eskola & Suoranta, 1998). Harkinnanvarainen otanta rajaa haastateltavia siten, että haastateltavilla on oletettavasti hyvä tuntemus käsiteltävästä aihepiiristä, mikä on laadulliselle tutkimukselle oleellista (Tuomi & Sarajärvi, 2018).

Tässä työssä keskeisenä seikkana oli se, että haastateltavat käsittelivät samaa tietopankkia. Haastateltavat koostuivat siis harkinnanvaraisesti valikoiduista potentiaalisen tietopankin käyttäjistä ja täydentäjistä. Haastateltavia valittaessa tavoitteena oli saada henkilöitä keskeisistä datatieteellisistä rooleista. Saltz ja Grady (2017) kartoittavat erinäisistä lähteistä selkeimmiksi rooleiksi datatieteilijän (*engl. data scientist*), data-arkkitehdin (*engl. data architect*), data-analyytikon (*engl. data analyst*) sekä datainsinöörin (*engl. data engineer*). Kohdeyrityksellä on edellä mainittujen roolien piiristä työntekijöinä data-arkkitehteja ja datainsinöörejä. Lisäksi yrityksestä löytyy myös koneoppimisinsinöörin

rooli, joka on datainsinöörin koneoppimiseen erikoistuva variantti. Yrityksestä saatavilla olevien roolien ja kartoitettujen datatieteellisten roolien perusteella harkinnanvarainen otanta kohdistui data-arkkitehteihin, datainsinööreihin ja koneoppimisinsinööreihin.

Samasta tietopankista keskusteleminen usean haastateltavan kanssa mahdollisti aiheen laajan käsittelyn. Haasteena oli toki se, että tietopankkia ei ollut olemassa, mutta siitä huolimatta tavoitteessa onnistuttiin. Taulukossa 1 on koostettu haastateltavat, heidän roolinsa, haastattelun painopiste sekä haastattelun kesto.

Taulukko 1. *Haastateltavat*

Nimi	**Rooli**	**Haastattelun painopiste**	**Haastattelun kesto**
Data-arkkitehti 1	Data-arkkitehti	Järjestelmän laatu	46 min
Data-arkkitehti 2	Data-arkkitehti	Käyttöaikomus ja käyttö	42 min
Data-arkkitehti 3	Data-arkkitehti	Käyttöaikomus ja käyttö	39 min
Data-arkkitehti 4	Data-arkkitehti	Järjestelmän laatu	51 min
Datainsinööri 1	Datainsinööri	Informaation laatu	42 min
Datainsinööri 2	Datainsinööri	Järjestelmän laatu	42 min
Datainsinööri 3	Datainsinööri	Järjestelmän laatu & käyttöaikomus ja käyttö	27 min
Datainsinööri 4	Data arkkitehti / data insinööri	Järjestelmän laatu	37 min
Koneoppimisinsinööri 1	Koneoppimisinsinööri	Informaation laatu	43 min
Koneoppimisinsinööri 2	Koneoppimisinsinööri	Nettohyödyt	41 min

Taulukossa 1 huomioitu painopiste syntyi haastattelun aikana luonnollisesti. Tavoitteita tiettyihin painotuksiin ei siis ennalta ollut.

4.4 Aineiston analysointi

Teemahaastattelujen johdattelemana aineistoa lähdettiin analysoimaan teemoittelun kautta. Taustalla tässä kuitenkin toimii laadulliselle tutkimukselle ominainen aineiston koodaus, jossa aineistosta konstruoidaan helpommin pilkottavia osia, jotka toimivat tutkijan tulkintoina tietyistä osista aineistoa (Eskola & Suoranta, 1998; Braun & Clarke, 2006). Kun aineisto on koodattu, se on vaivattomampi viedä teemoittelun äärelle. Tässä työssä koodaus tehtiin täysin aineistolähtöisesti ja myöhemmin sidottiin teoriaan teemojen kautta.

Vielä kuitenkin ennen koodausta, aineisto muunnettiin sitä tukevaan muotoon. Tässä tapauksessa se oli kombinaatio kahdesta asiasta. Ensiksi nauhoitetut haastattelut vietiin automaattisen litteroinnin läpi, jolloin saatiin karkeat vedokset haastatteluista tekstimuodossa. Vedoksissa kuitenkin esiintyi jonkin verran virheitä etenkin englanninkielisten termien kanssa, mutta pääasiassa vedokset olivat ymmärrettäviä. Virheitä ei sen enempää korjattu, koska ne tulevat selvennetyksi kombinaation toisen osapuolen kanssa. Tämä toinen osa koodausta tukevasta muodosta ovat itse tallenteet haastatteluista. Tallenteet vietiin karkeiden litterointien kanssa Atlas.ti -ohjelmistoon, joka mahdollistaa niiden sujuvan rinnakkaistarkastelun haastattelun edetessä. Tarkemmin sanottuna tämä mahdollistaa koodien asettamisen tekstimuotoiseen tiedostoon juuri siihen kohtaan ja ajanhetkeen, jossa toistossa oleva tallenne menee. Litteroinnissa aiemmin esiintyneet virheet eivät siis ole haitaksi, koska haastatteluiden tallenteet ovat niitä aina tukemassa.

Koodattu aineisto jaoteltiin edelleen Atlas.ti -ohjelmiston avulla alaluvussa 2.4 esitellyn DeLonen ja McLeanin (2003) viitekehyksen teemoihin. Näitä ovat informaation laatu, järjestelmän laatu, palvelun laatu, käyttöaikomus ja käyttö, käyttäjätyytyväisyys ja nettohyödyt. Tällaista teoriasidonnaisuutta hyödyntämällä voidaan Tuomen ja Sarajärven (2018) sekä Puusan et al. (2020) mukaan tehdä uusia havaintoja ja löytää uusia näkökulmia tarkasteltavista ilmiöistä. Teoreettinen viitekehys tarjoaa oivan lähtökohdan aineiston tulkinnalle ja analyysille, mutta itse analyysin tulokset voivat kuitenkin olla laajempia ja tuottaa uusia teoreettisia näkökulmia. Tutkimus täten mahdollistaa myös uusien tulkintojen ja johtopäätösten syntymistä eikä ainoastaan vahvista olemassa olevaa teoreettista viitekehystä, kuten puhtaalla teorialähtöisellä analyysimenetelmällä on tapana. Teemojen sisältöä ja analyysin tuloksia tarkastellaan seuraavassa luvussa tarkemmin.

5. TULOKSET

Aineiston analysoinnin tuottamia tuloksia tarkastellaan informaation laadun, järjestelmän laadun, palvelun laadun, käyttöaikomuksen ja käytön, käyttäjätyytyväisyyden ja nettohyötyjen kokonaisuuksien kautta. Lisäksi lopussa koostetaan kaavio, jossa edellä mainittujen kokonaisuuksien keskeisimmät tulokset ja kokonaisuuksien vuorovaikutus ovat havaittavissa. Koska tiedon eri lajit ja konversio ovat näkyviä aiheita tietopankeista puhuttaessa, tulosten tarkastelussa tuodaan myös esiin havaintoja, jos esimerkiksi jokin järjestelmän laadullinen tekijä tukee vahvasti jotain tiettyä SECI-mallin vaihetta ja mitä se käytännössä tarkoittaa.

Taustaksi ja tutkimusongelman vahvistamiseksi liittyen aineistosta havainnoitiin, että yleisesti ottaen haastateltavat kertoivat kokevansa toistettavuutta datatieteellisissä projekteissa. Toistettavuutta koettiin esimerkiksi kuvassa 9 esiintyvien datalähteiden tarkastelun kanssa, mikä tarkemmin sanottuna kuvaa tapahtumaa, jossa pohditaan millä keinoin tietynlaisesta lähteestä tuodaan dataa varastointiin. Lisäksi tietomallien luontia, nimeämiskäytänteitä, teknologioiden valintaa ja mallien arviointia pidettiin toistuvina tehtävinä. Toistettavia asioita on siis rutkasti ja mikä tärkeintä, niitä esiintyy läpi datatieteellisten projektien toteutuksien. Tässä luvussa analysoidaan aineistoa ja luodaan sitä kautta kehys tietopankille toistettavuusongelmien avuksi. Braunin ja Clarken (2006) havaitsemien hyvien teemoittelun käytänteiden mukaisesti tulosten tarkastelussa ei pelkästään tyydytä aineiston esittelyyn, vaan siitä pyritään koostamaan johtopäätöksiä, jotka ovat tutkimuskysymyksille relevantteja.

5.1 Informaation laatu

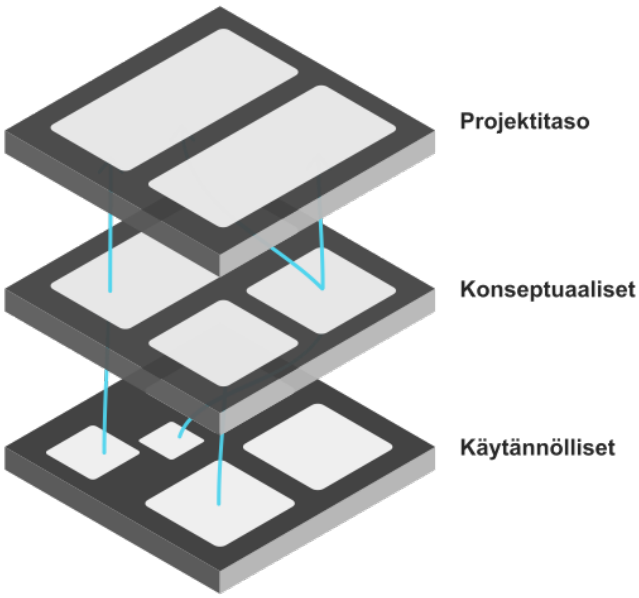
Ennen kuin voidaan tarkastella informaation laatuun liittyviä tekijöitä, tulee tarkastella itse informaatiota, jota tietopankista tulisi tulosten perusteella löytyä ja lisäksi missä muodossa varastoitu informaatio on. Tietopankin sisältö on keskiössä, sillä tarkastelussa olevalle tietopankille on selkeä rajattu käyttötarkoitus datatieteellisten projektien tukena. Aineiston perusteella sisältö voidaan rajata kolmeen eri tasoon: projektitaso, konseptuaalinen ja käytännöllinen. Taustalla on haastateltavien toiveet sisällöstä, kuten konkreettisista esimerkeistä, infrastruktuureista ja arkkitehtuureista sekä projekteissa toimineista kokonaisuuksista.

Konseptuaalisella tasolla tarkoitetaan tässä yhteydessä ylätason asioita kuten menetelmiä, teorioita, suunnitteluperiaatteita ja toimintatapoja, jotka ohjaavat datatieteellisten

projektien infrastruktuurillisia valintoja. Tällä tasolla pyritään tarjoamaan tietoa esimerkiksi siitä, millainen infrastruktuuri tukee projektia, jossa datalähteenä toimivat useat tietokannat ja jossa liiketoiminta tavoittelee raportointia. On syytä tarkentaa, että tämä konseptuaalinen taso käsittelee datatieteellisen projektin infrastruktuurillisia kokonaisuuksia, kuten datalähteitä, varastointia ja lopputuotteita erikseen omina kokonaisuuksinaan.

Tieto siitä mitkä variaatiot kokonaisuuksista toimivat keskenään on ylimmän eli projektitason sisältöä. Kun näihin on liitetty perustelut, kokemukset sekä hyvät ja huonot puolet, tietopankin käyttäjä kykenee vertailemaan oman projektinsa liiketoiminnan tarpeita ja lähtökohtia ja sitä kautta pohtimaan sopivia infrastruktuuriratkaisuja. Tämän SECI-mallin mukaisen sisäistämisvaiheen kautta tietopankista löytyvää tietoa päästään hyödyntämään uudelleen toisaalla.

Alin taso sisällöstä kattaisi puolestaan keinot aiemmin mainittujen infrastruktuurillisten valintojen toteutukseen käytännössä. Niihin lukeutuu muun muassa haastateltavien suosimat infrastructure-as-code (IaC) sapluunat, joiden tarkoituksena on perustaa infrastruktuurista löytyvät komponentit tarvittavilla teknologioilla automaattisesti. Sapluunoiden lisäksi käytännöllisen sisältötason tulisi sisältää eri ohjelmointikielillä kirjoitettuja geneerisiä funktioita ja moduuleita, jotka ovat tukemassa konseptuaalisen tason kokonaisuuksia. Etenkin nämä käytännön tason asiat ovat keskeisiä, koska ne tarjoavat pelkän dokumentaation lisäksi funktionaalista sisältöä. Mainittuja sisältötasoja ja niiden linkittymisiä on havainnollistettu seuraavalla kuvalla.



Kuva 11. *Tietopankin sisältötasot*

Kuvan 11 mukaisen usean tason jaottelun rikkaus on se, että eri konseptuaalisia ratkaisuja voidaan käyttää ristiin ja sisällyttää useissa eri projektikokonaisuuksissa. Lisäksi

tarjoamalla näihin käytännön toteutukset, datatieteellisten projektien iteratiivinen eteneminen on sujuvampaa.

Esitellyn sisällön laatuun vaikuttavat etenkin sen ajankohtaisuus ja oikeanlaisuus. Ajankohtaisuuden varmistamiseksi esitetään ehdotuksia sitä tukevista ominaisuuksista seuraavassa alaluvussa. Oikeanlaisesta sisällöstä puolestaan pohdittiin haastatteluissa muun muassa seuraavaa:

”Kyllä mulla itselläni on joku visio siitä, että mitä sinne haluaisin laittaa. Mutta kun en mä tiedä mitä kaikkea muut sinne haluaisivat. … muiden käyttäjien odotukset ja sitten mun mielikuva siitä, että mitä mä sinne haluaisin laittaa pitäisi saada kohtaamaan, jotta se sisältö olisi järkevää.”- Data-arkkitehti 4

Tästä pääteltynä oikeanlaista tietoa saavutetaan sillä, että tietopankkiin tallennettava sisältö on ennalta määritelty useamman osallistujan kanssa. Tietopankin alkuvaiheessa tulisi siis määritellä konkreettisesti ja suhteellisen tarkasti mitä sisältöä kullekin tasolle tallennetaan. Tietysti jo aiemmin esitetyt esimerkit sapluunoista ja infrastruktuurillisista ratkaisuista pitävät edelleen paikkansa, mutta tässä painotetaan tarkentamaan näitä.

5.2 Järjestelmän laatu

Järjestelmän laatu kattaa tässä tapauksessa tietopankin tavoiteltuja teknisiä ominaisuuksia. Toivottuja ominaisuuksia nousi haastatteluista paljon esiin ja tässä alaluvussa koostetaan niistä keskeisimmät ja yleisimmät. Ominaisuuksia kartoitettiin ajatuksella, että tietopankki olisi haastateltavien työn tukena usean eri asiakkaan kanssa ja siten myös useassa datatieteellisessä projektissa, mahdollisesti myös samanaikaisesti.

Ominaisuuksia kysyttäessä, hakutoiminto oli lähes poikkeuksetta ensimmäinen vastaus. Tämä ei yllätä, sillä järjestelmästä, jonka tehtävänä on varastoida ja pitää sisällään organisaatiolle keskeistä tietoa, toivoo löytävänsä helposti etsimänsä. Hakutoiminnon kanssa on huomioitava ensinnäkin, että se kohdistuu otsikkotasolta myös sisältötasolle. Näin parannetaan mahdollisuuksia oikean tiedon löytämiseksi. Toinen huomioitava asia haun kanssa on termistö. Haastateltavat nostivat huolensa tilanteista, joissa haku ei onnistu löytämään haluttua tietoa yhdyssanavirheiden, kirjoitusvirheiden tai samaa tarkoittavien sanojen takia. Näiden huomioitujen asioiden perusteella hakutoiminnon suunnitteluun tulee esimerkiksi ottaa käyttöön laaja metatieto ja muu kategorisointi, joihin haku voi tukeutua.

Aineiston perusteella tietopankissa on syytä pitää kirjaa tallennetun tiedon omistajuudesta ja keskeisitä ajankohdista kuten tallennushetkestä, muokkaushetkestä ja uudelleenkäyttöhetkestä. Omistajuus koetaan aineiston perusteella vaikuttavan yllättävänkin

moneen asiaan. Se kasvattaa tallennetun tiedon luotettavuutta, helpottaa yksilötasolla uuden tiedon tallentamisen seurantaa ja madaltaa kynnystä lisäselvitykseen. Tiedot keskeisistä ajankohdista puolestaan selventävät ja vahvistavat tallennetun tiedon ajankohtaisuuden. Näiden aikaleimojen perusteella pystyy olemaan omistajaan yhteydessä uudelleenarviointia varten, mikäli tieto on selkeästi vanhentunutta. Tiedon vanheneminen on koettu hyvin suureksi haasteeksi tietopankin ääreltä, minkä vuoksi siihen tulee keskittyä teknisessä toteutuksessa. Yhteydenotto voi parhaassa tapauksessa olla myös automaattinen prosessi, joka käynnistyy tietyillä ajanhetkillä.

Vaikka omistajuus kuulostaa siltä, että se sitoo sisällön aina yhteen ihmiseen, näin ei ole. Kaikesta yksilösidonnaisuudesta pyritään tietopankilla pääsemään eroon. Tätä tukee haastateltavien toiveet ja näkemykset vahvasta kollaboraatiosta tietopankin äärellä. Järjestelmästä on löydyttävä ominaisuus, joka mahdollistaa yhteistyön tiedon lisäämisvaiheessa ja myöhemmin muokkaamisessa. Tähän liittyen pitää olla mahdollisuus hyväksyä tai pyytää muutoksia, mikä puolestaan luo vaatimuksen vertaisarvioinnin ominaisuudesta.

Kollaboraatio-ominaisuus itseasiassa koostuu perusperiaatteiltaan SECI-mallin ulkoistaminen ja yhdistely vaiheista. Kollaboraatiossa useat yksilöt eksplikoivat hiljaista tietoa toisten ymmärrettäviksi käsitteiksi ja malleiksi (ulkoistaminen), ja useasta yksilöstä koostuva ryhmä tulee lopulta konsensukseen aiheesta ja yhdistää tiedon aiemmin tallennettuun sisältöön tietopankissa (yhdistely). Näin jo tässäkin vaiheessa tietopankin käyttöä kyetään luomaan uutta tietoa osallistujien kesken.

Jo mainittu yhteistyö tiedon lisäämisvaiheessa vaatii aineiston mukaan tietopankilta useat eri laatutasot. Sisältöä voidaan esimerkiksi työstää alimmalla tasolla ja yhdistellä olemassa olevaan tietoon sitten seuraavalla tasolla. Hakutoimintoa puolestaan voi kohdistaa vain ylimmän tason yhdisteltyyn ja vertaisarvioituun sisältöön. Alimmalla tasolla työstämisestä ja kollaboraatiosta päästäänkin seuraavaan tekniseen ominaisuuteen – versionhallinta. Samaan tapaan kuin ohjelmistokehityksessä, tietopankkiin tallennettavaa sisältöä, oli se sitten koodia tai dokumentoituja suunnitteluperiaatteita, pitää pystyä jatkokehittämään versiosta toiseen ja tutkia versiohistoriaa. Käytännössä versionhallinta on se, joka mahdollistaa sujuvan kollaboraation.

Tallennettua tietoa pitää kyetä käyttäjän tai automaation toimesta viittaamaan johonkin muuhun tallennettuun tietoon. Tämäntapaiset linkitykset ovat yksi varsinaisista käytännön keinoista, joilla informaation laatu -luvussa havainnollistettu sisällön tasoerittely toteutetaan. Viittausten ja linkityksien kautta tiedon etsijää ohjataan relevantin sisällön luokse.

Koska tietopankin sisältö ei voi aina vastata täysin tarpeisiin, tulisi siinä olla ominaisuus, jonka kautta toivotusta – tällä hetkellä puuttuvasta – sisällöstä voisi tehdä niin sanotusti virallisen pyynnön. Tämän kautta toive tulisi kirjattua ja näin ollen helpommin huomioitua tulevaisuudessa. Yhtä lailla on mahdollista, että joku huomaa toiveen ja ohjaa etsijän sellaisen sisällön pariin, joka vastaa toivottua.

Jotta tietopankki olisi yhtenäinen, tallennusvaiheessa tulisi olla tietty sapluuna, jonka mukaisesti uutta tietoa tallennetaan. Yhtenäisyys helpottaa etenkin tiedon uudelleenkäyttöä, kun keskeinen sisältö löytyy aina samasta paikasta samalla tavalla. Lisäksi jo mainittu kollaboraatio on selkeämpää, kun työskentely käydään aina samanmuotoisen sisällön ääreltä.

Yrityksillä on yhä enemmän erilaisia tietojärjestelmiä käytössään, joista jokainen palvelee tiettyjä spesifejä yrityksen osa-alueita. Järjestelmien muodostavan kompleksisen kokonaisuuden vuoksi eräs haastateltava nosti tietopankkeja ajatellen yhdeksi keskeiseksi tekijäksi sen, että työntekijät läpi organisaation olisivat tietoisia sieltä löytyvästä tiedosta.

”Tiedon löytyminen on yrityksissä todella vaikeaa. Se on aina piilotettu kaikenlaisiin erilaisiin verkkojärjestelmiin ja se että osaa etsiä oikeasta paikasta oikealla tavalla, tekee löytämisestä todella vaikeaa. Niin kaikki mahdolliset promoamiset auttaisivat siinä.” - Data-arkkitehti 1

Hän tarkentaa ehdottamaansa promoamisominaisuutta, että kaiken promoamisen tai mainostamisen ei tulisi olla tilattua, sillä etenkin uudet työntekijät eivät välttämättä edes tiedä mitä etsiä. Tarvitsee siis niin sanotusti tyrkyttää tietopankista löytyvää tietoa ajoittain tarjolle esimerkiksi erilaisten viestintäpalveluiden kautta. Promoaminen on yksi monista asioista, joilla pidetään tietopankki organisaation työntekijöiden tietoisuudessa ja siten ohjataan käyttöä siihen. Promoaminen ja muut haastatteluista kerätyt keskeisimmät ominaisuudet koostettiin lyhyiden selityksien kanssa seuraavaan taulukkoon.

Taulukko 2. *Tietopankin tekniset ominaisuudet*

Ominaisuus	**Selitys**
Hakutoiminto	Tarjoaa vapaata tekstikenttää vastaan tietopankista relevanttia sisältöä.
Kategorisointi	Sisällön ryhmittely metadatan yms. avulla joko automaattisesti tai manuaalisesti.
Omistajuus	Kertoo kuka/ketkä ovat tarjonneet tiedon tietopankkiin.

Ajankohtaisuuden seuranta	Aikaleimoihin perustuva tukitoiminto, joka huolehtii, ettei tietopankin sisältö vanhene.
Kollaboraatio	Mahdollistaa usean henkilön samanaikaisen työskentelyn sisällön äärellä.
Vertaisarviointi	Varmistaa, että tallennettava sisältö vastaa odotettua laatua ja useamman henkilön näkemyksiä.
Laatutasot	Mahdollistavat mm. haun tapahtumisen ylimmällä laatutasolla ja kollaboraation tapahtumisen sitä alemmalla laatutasolla.
Versionhallinta	Tarjoaa historiatiedot sisältömuutoksista ja sujuvan jatkokehityksen.
Viittaukset/linkitykset	Yhdistelee mm. eri sisältötasojen tietoa toisiinsa.
Sisältöpyynnöt	Ominaisuus, jonka avulla voi tehdä pyynnön tietopankkiin tallennettavasta sisällöstä.
Tallennusvaiheen sapluuna	Koostuu ennalta määritetyistä kentistä ja pitää tietopankin sisällön yhtenäisenä.
Promoaminen	Mainostaa ja nostaa esiin tietopankin sisältöä esimerkiksi viestikanavissa.

Taulukossa 2 listatut tekniset ominaisuudet antavat suuntaviivat tietopankin varsinaista teknistä toteutusta varten. Keskeiset järjestelmän laadulliset tekijät ovat yhdistetty muihin tietojärjestelmien onnistumisen teemoihin myöhemmin alaluvussa 5.7.

Haastateltavat pitivät myös tietopankin integraatiokyvykkyyksiä muihin tietojärjestelmiin tärkeänä. Taustalla on lisätyön minimointi ja tietopankin käytön tuominen mahdollisimman lähelle muuta työskentelyä sujuvan käyttökokemuksen aikaansaamiseksi. Integraatio on tärkeä myös jatkokehitystä ajatellen. Esimerkiksi uudenlaisten tekoälyratkaisujen tuominen tukemaan tiedon etsimistä tietopankista vaatii siltä kykyä integroitua jollakin tasolla.

5.3 Palvelun laatu

Palvelun laadun ja tarvittavan käytön tuen varmistamiseksi ehdotettiin tietopankille fasilitaattoria, jonka tehtävänä olisi varmistaa muun muassa tallennetun tiedon ajantasaisuus. Fasilitaattorin tai vastaavan vastuutahon esittäminen tietopankille voisi yhtä lailla tarkoittaa vastuutahon nimeämistä koko tietämyksenhallinnalle.

Vaikka tietty vastuutaho tulisikin nimettyä, tietämyksenhallinta on lopulta kuitenkin yhteisen tekemisen takana. Fasilitaattori auttaa ennen muuta saavuttamaan tietämyksenhallinnalle ja tietopankille sellaisen tason, että sitä voidaan pitää yhteisenä tekemisenä. Lisäksi fasilitaattori voitaisiin nähdä selkeänä tahona, joka on vastuussa tietämyksenhallinnan aloitteiden, kuten tietopankin, jatkokehittämisestä.

Palvelun laatuun liittyy myös tietopankin saavutettavuus. Aineistosta on havaittavissa toive tai oikeastaan jopa vaatimus, että tietopankin tulee kytkeytyä saumattomasti jokapäiväiseen tekemiseen. Tietopankki ei saa olla niin sanotusti monen klikkauksen päässä. Jos tässä onnistutaan, varmistetaan alhaisempi kynnys käytölle. Hyvä saavutettavuus pätee sekä tiedon tallentamiseen että tiedon hakemiseen.

5.4 Käyttöaikomus ja käyttö

Käyttöaikomus ja käyttö käsittävät tietopankin kontekstissa tiedon tallentamista, tiedon muokkaamista, tiedon vertaisarviointia ja tiedon hyödyntämistä. Kaikkia näitä näkökulmia pyritään tarkastelemaan tässä alaluvussa.

Delonen & McLeanin (2003) viitekehyksen mukaisesti käyttöaikomukseen ja käyttöön vaikuttavat edellisissä alaluvuissa esitellyt informaation laatu, järjestelmän laatu sekä palvelun laatu. Näiden lisäksi haastatteluista ilmeni myös kulttuurillinen näkökulma. Koska tietopankki on lopulta vain teknologioilla mahdollistettu tietämyksenhallinnan väline, tietämyksenhallinnan toteuttaminen organisaatiossa vaatii kulttuurillista muutosta. Ilmaisutavat haastattelijoilla vaihtelivat, mutta ajatukset laaja-alaisen tietopankin käytön onnistumisesta vain ”luonnollisin” keinoin pysyivät samana.

”Jos sanoisin, että hei, meillä on nyt tällainen tietopankki, tulkaa kaikki lisäämään, niin ei tule. Mutta sitten kun et kerro siitä kenellekään, vaan näytät yhdelle ihmiselle, että hei, meillä on täällä tällaista, niin se tulee innoissaan mukaan lisäämään ja se kertoo jossain omalle kollegalleen. Tavallaan sitä kautta se käyttö tulee orgaanisesti tietopankille…” - Datainsinööri 4

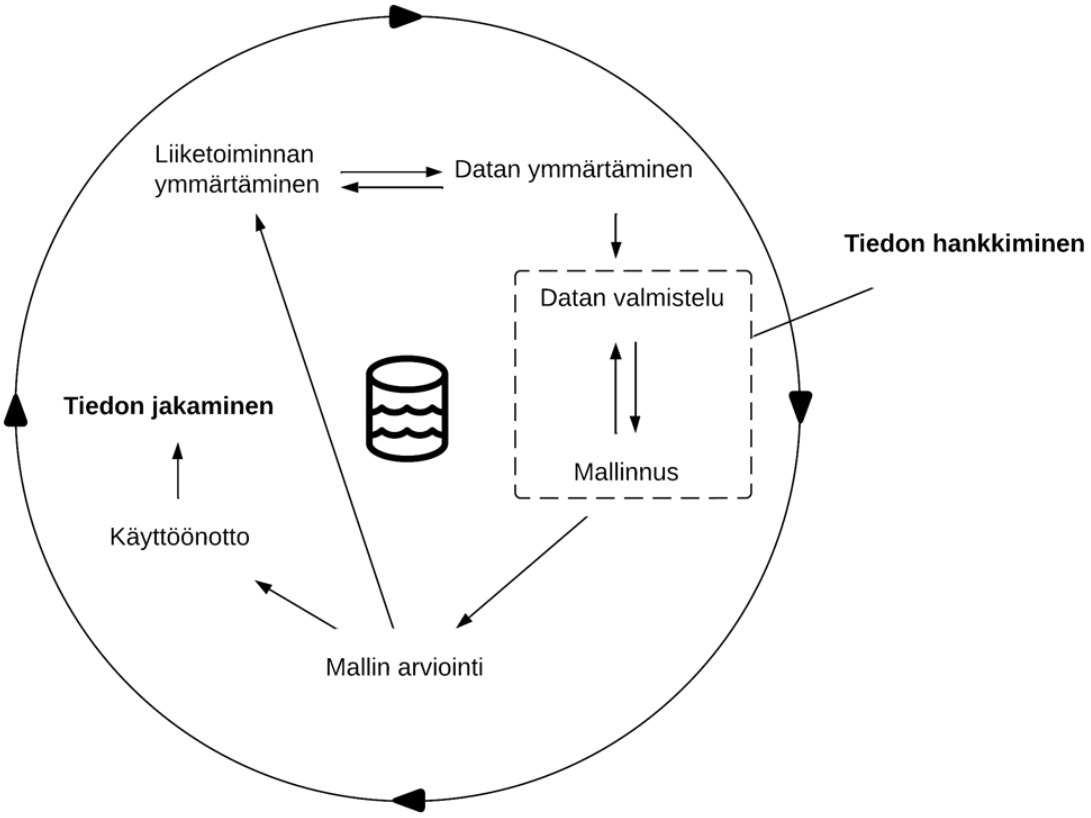
Edellä olevassa Datainsinööri 4:n kuvailussa on havaittavissa loistava idea, miten luonnollinen käyttöönotto ja kulttuurillinen muutos on mahdollista saada aikaan. Varsinainen

käyttöönotto tulisi tehdä siis mieluummin hallitusti pieni ryhmä kerrallaan kuin heti laajan organisaatiotason lanseerauksen kanssa. Tämä on keskeinen asia, kun tietopankkia pyritään ottamaan laajemmin käyttöön, mutta tässä työssä ei paneuduta kyseiseen aihepiiriin enempää.

Selkeästi käyttöaikomusta ja käyttöä laskee, jos tietopankista ei löydy tietoa. Tiedon löytymättömyydelle aineiston mukaan voi olla muutama selkeä selitys, joihin kaikkiin pitää tarttua tietopankkia suunniteltaessa. Ensinnäkin tietoa ei välttämättä ole saatavilla, mitä voidaan pyrkiä välttämään tietopankin alkuvaiheen minimisisällöllä ja myöhemmin pyynnöillä sisällöstä. Toiseksi, kun tiedon määrä kasvaa, sen etsiminen vaikeutuu, mikä korostaa aiemmin työssä nostettua hakutoimintoa entisestään. Aiheutuva käytön lasku luonnollisesti laskee tietopankista saatavia hyötyjä niin organisaatio- kuin yksilötasolla.

Tiedon löytyminen on tärkeää käytön kannalta, mutta niin on myös tiedon tallentamisen helppous. Tämä nostettiin yhtenä keskeisenä käyttöön ja käyttöaikomukseen vaikuttavana tekijä. Jos se ei ole helppoa ja vaivatonta, uuden tiedon tallentaminen normaalin työnteon ohessa voi koitua liian raskaaksi ja aikaa vieväksi prosessiksi. Jos päädytään tällaiseen tilanteeseen, jossa uutta tietoa ei tallenneta tietopankkiin, sen käyttö hiipuu samaa tahtia kuin sen sisältö. Välttääkseen kuvailtua tilanteita, tallennusvaiheen helppokäyttöisyys ja matala kynnys ovat elintärkeitä. Tässä sisällön laatutasot koetaan ainakin osittaisena ratkaisuna. Alin taso keskitetyssä tietopankissa olisi sujuvasti saavutettava paikka hyvin alkuvaiheen luonnostelmille eksplikoidusta tiedosta, joihin pystyisi myöhemmin palata ja viemään tietämyksenhallinnan prosessia eteenpäin yksin tai yhdessä.

Aineistosta voidaan tulkita, että tietopankin käyttö ei saa tuntua niin sanotusti lisätyöltä jo olemassa olevan työn rinnalla. Varsinkin sellainen tietopankki, jolla on selkeä ja rajattu käyttötarkoitus pitää saada osaksi työnteon arkea ja projektien tekemistä. Tähän ehdotetaan seuraavassa kuvassa esitettyä tietopankin käytöllä rikastettua datatieteellistä prosessia.



Kuva 12. *Tietopankin käyttö osana datatieteellistä prosessia*

Ennen tietopankin käyttämistä, datatieteellisen projektin on edettävä liiketoiminnan ymmärtämisen ja datan ymmärtämisen vaiheiden läpi, sillä näistä vaiheista datatieteilijä rakentaa käsityksen projektin ennakkoasetelmista. Siirryttäessä datan ymmärtämisestä datan valmisteluun, tietopankki esitellään prosessiin kuvasta 12 löytyvällä *Tiedon hankkiminen* -kokonaisuudella. Tavoitteena tässä on datan valmistelun ja mallinnuksen vaiheiden aikana hankkia tietopankista ohjeita ja parhaita käytänteitä miten jotain tietynlaista projektia tulisi edistää, mitkä infrastruktuuriratkaisut toimivat tilanteessa parhaiten ja mitkä mallinnusmenetelmät sopisivat. Optimaalisessa tapauksessa tietopankista löytyy alimmalta sisältötasolta käytännöllinen ratkaisu käsillä olevaan ongelmaan. Sen lisäksi, että prosessimallia suorittava datatieteilijä saa ison hyödyn tukeutuessaan muiden kokemuksien kautta saatuun tietoon, tulee hänen ymmärtää palautteen annon tärkeys. Tiedon hankkimisen jälkeen kaikki tarkennukset ja kommentoinnit jalostavat sisältöä entisestään, mikä voi jopa johtaa alkuperäisen tiedon tallentajan tekemään muutoksia omassa projektissaan.

Myöhemmin, kun mallit sekä menetelmät ovat valittu, arvioitu ja käyttöönotettu, tietopankin käyttö tulee jälleen prosessimallissa esille. Koska jokainen projekti on kuitenkin uniikki, tieto mitä konsepteja – ja ennen kaikkea miksi – esimerkiksi datan tuonnin, va-

rastoinnin ja transformaation kanssa juuri tässä projektissa käytettiin, tulee jakaa ja tallentaa tietopankkiin. Laajempaa ja monipuolisempaa sisältöä tietopankissa tulee siten kerrytettyä. Kun seuraavan projekti taas etenee datan valmisteluun ja mallinnukseen, kattavampaa materiaalia on oletettavasti saatavilla työn tueksi.

Tämän tapaustutkimuksen tulosten perusteella ehdotetaan muokattua datatieteellistä prosessimallia, jotta tietopankin käyttö on todellisuudessa mahdollisimman todennäköistä. Pelkästään se, että tietopankista tiedon hankkiminen on kirjattuna välivaiheena laajemmassa prosessissa kasvattaa isomman yleisön tietoisuutta asiasta ja vie tietopankin käyttöä osaksi arkea, vaikka kaikki eivät jokaisessa projektissa sitä käyttäisikään.

Tietopankin käytön osana datatieteellistä prosessia voi yhdistää vielä laajempaankin kokonaisuuteen. Aiemmin työssä esitetyissä kuvissa 2 ja 4, joissa havainnollistetaan tietämyksenhallintaa osana liiketoimintaprosesseja ja/tai projekteja, on vaiheet *Tiedon jakaminen* sekä *Tiedon hankkiminen*. Näitä vaiheita voidaan pitää samoina vastaavina vaiheina, jotka löytyvät kuvasta 12.

5.5 Käyttäjätyytyväisyys

Käyttäjätyytyväisyyttä ja samalla motivaatiota tietopankin käyttöön pitää tarkastella tiedon hankkimisen ja tiedon jakamisen kannalta, sillä kokonaistyytyväisyyteen vaikuttaa kokemukset kummastakin tapahtumasta. Positiivista oli, kun haastatteluissa mainittiin useaan otteeseen, että muiden saama hyöty motivoi tallentamaan jatkossakin lisää uutta tietoa tietopankkiin.

Tähän liittyen pohdittiin tarpeellista ominaisuutta, jonka avulla saataisiin tiedon omistajalle ilmoituksia tai viestejä, että lisäämänne tietoa on käytetty kuvan 12 tiedon hankkiminen -vaiheen kautta ja siitä on ollut apua. Eli kun koettu hyöty saadaan viestittyä osallisille, käyttäjätyytyväisyys saadaan nousuun ja positiivinen kierre alkaa ja tehostaa tietopankin käyttöä.

Tiedon hankkimisen osalta käyttäjät pysyvät haastattelujen mukaan tyytyväisenä, kun käyttö koetaan hyödylliseksi oman työn tukena. Konkreettisesti hyödyn kokeminen kumpuaa laadukkaasta, ajankohtaisesta ja helposti saavutettavasta sisällöstä, joka auttaa ratkomaan datatieteellisen projektin ongelmia. Edellä mainitut asiat eivät ole yksinkertaisesti tavoitettavissa, mutta niiden edistämistä helpottaa muun muassa aiemmin mainitut järjestelmän ominaisuudet.

5.6 Nettohyödyt

Tietopankin käyttämisen kautta saavutettujen hyötyjen taustalla on asiantuntevien ihmisten kokemukset ja tietämys saatettuna eksplisiittiseen muotoon muiden käytettäväksi. Keskeiset toimintatavat ja menetelmät siirtyvät yksilöihin sitoutuneesta hiljaisesta tiedosta paremmin saavutettavaksi. Hyödyt datatieteellisten projektien tukena olevasta tietopankista voidaan aineiston perusteella jaotella karkeasti kahteen osaa. Hyödyt, joita organisaatio saavuttaa ja hyödyt, joita yksittäinen käyttäjä saavuttaa. Näissä molemmissa osissa vaikuttaa haastateltavien näkemys, että ideaalitapauksessa tietopankin käyttö säästää aikaa ja siten työaikaa voi kohdistaa tehokkaammin.

Organisaatiotason hyödyistä haastateltavat kokivat yhteisten työtapojen selkeytymisen keskeisenä. Tallennetun tiedon yhtenäisyys ja yhdenmukaisuus syntyy ennalta määritellyn tallennusvaiheen sapluunan sekä työstövaiheen kollaboraation kautta. Kehitetyt yhtenäiset työtavat näyttäytyvät tasaisen laadukkaana työskentelynä asiakkaasta ja projektista toiseen ja siten luovat hyvää kuvaa asiantuntijaorganisaatiosta. Tietopankin avulla organisaatiolle muodostuu kollektiivista osaamista, jonka vuoksi asiakkaita voidaan palvella nopeammin ja laadukkaammin.

Tietopankki mahdollistaa myös organisaatiolle uusien työntekijöiden sujuvamman perehdyttämisen. Eksplisiittiseen tietoon on yksinkertaista tukeutua ja sen avulla esitellä miten juuri tässä organisaatiossa tehdään asiat ja toimitaan datatieteellisissä projekteissa. Näin saatetaan uusi henkilö yhteisten työtapojen äärelle ja saavutetaan edellä mainitut hyödyt.

Yksilötasolla tietopankin käytön nähdään toimivan oppimisen tukena ja välineenä kouluttautumisessa. Parhaassa tapauksessa uuden asian purkaminen lähtee liikkeelle ylimmältä eli projektitasolta ja syvenee siitä alaspäin. Yksilöllä on siis tilaisuus SECI-mallin mukaisesti sisäistää eksplikoitua tietoa.

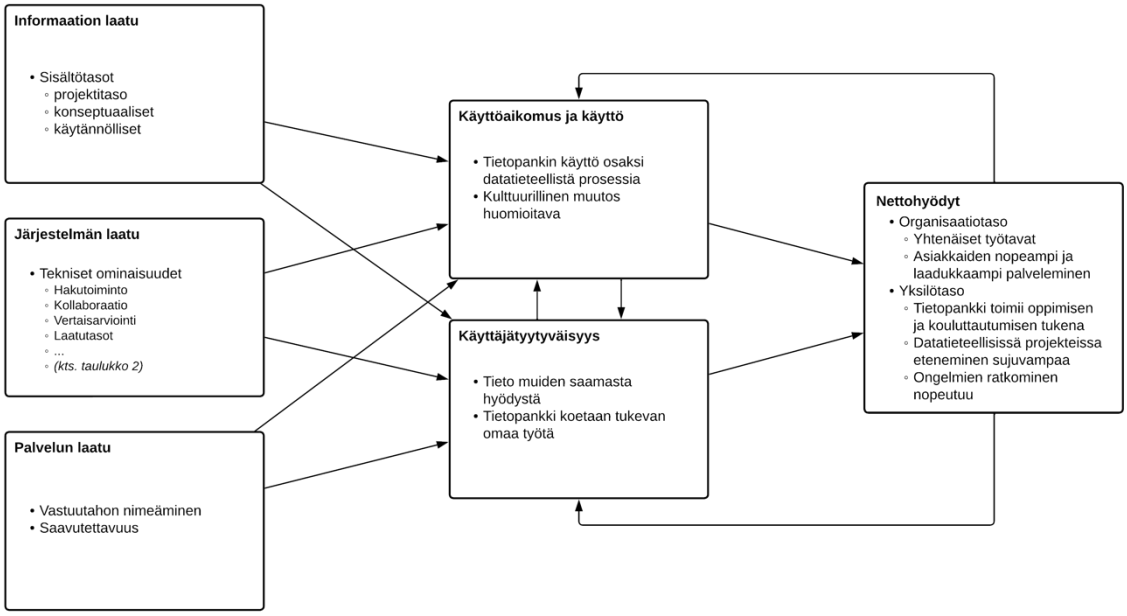
Datatieteellisten projektien kontekstissa yksilöt hyötyvät, kun pääsevät valmiiseen vertaisarvioituun tietoon käsiksi. Valmis tieto auttaa ennen kaikkea lähtemään uudessa datatieteellisessä projektissa liikkeelle, kun sieltä löytyy pohja mihin tukeutua. Toki myöhemmässä vaiheessa projektia voi ongelmatapausten ilmetessä tutkia tietopankkia esimerkiksi selvittämällä, miten jossakin konseptuaalisessa kokonaisuudessa ratkaistaan mallin arviointiin liittyvä ongelma. Sisällön tasojen takia, tähän ratkaisuun voi saada suoraan toimivan rivin koodia tai sapluunan tarvittavista resursseista hyvinkin nopeasti.

Lisäksi tietopankin ja tallennetun tiedon omistajuuden kautta yksilöiden kynnys lisäselvitykselle katsotaan madaltuvan. Tässä kuvailtu lisäselvitys tarkoittaa käytännössä SECI-mallin sosialisaatiovaihetta, jossa hiljaista tietoa välittyy yksilöiden välillä. Lisäselvitys on

otollista, koska se luo tilanteen, jossa selvityksen aluille laittanut henkilö voi jälleen ulkoistaa saamaansa tietoa eksplisiittiseen muotoon ja jakaa sitä takaisin tietopankkia.

5.7 Yhteenveto tuloksista

Tässä luvussa tarkastellaan kokoavalla otteella havaitut informaation laadun, järjestelmän laadun, palvelun laadun, käyttäjätyytyväisyyden, käyttöaikomuksen/käytön ja nettohyötyjen taustalla vaikuttavia tekijöitä. Keskeisimmät tekijät ovat koostettuna seuraavaan kuvaan työssä käytetyn Delone & McLeanin (2003) mallin rungon ympärille.



Kuva 13. *Yhteenveto tuloksista*

Kaikkiin kuvan 13 kokonaisuuksiin löytyi aineistosta näkökulmia. Informaation laatu, järjestelmän laatu, käyttöaikomus & käyttö ja nettohyödyt olivat kaikista vahvimmin esillä tuloksissa. Palvelun laatu ja käyttäjätyytyväisyys sen sijaan jäivät vähemmälle tarkastelulle.

Tietopankilta odotetaan käyttökelpoista sisältöä datatieteellisten projektien työskentelyn tueksi. Mikäli näihin odotuksiin ei kyetä vastaamaan, informaation laatu on yksi keskeinen epäonnistuja. Informaation laadun osalta datatieteellisten projektien tukena olevalle tietopankille ehdotettiin aineiston perusteella kolmijakoa: projektitaso, konseptuaaliset ja käytännölliset. Projektitaso kattaa tietyn projektin sisältämät konseptuaaliset valinnat eri infrastruktuuriratkaisujen taustalla. Konseptuaaliset valinnat kuvailevat puolestaan tarkemmin jonkin infrastruktuurin osan, kuten datan tuonnin tai transformaation menetelmiä. Käytännöllisen tason sisältö tarjoaa konseptuaalisille kokonaisuuksille toteuttamiskelpoiset resurssit.

Järjestelmän laadun alle tavoiteltuja teknisiä ominaisuuksia kerättiin kiitettävästi. Näihin kuuluu muun muassa hakutoiminto, kollaboraatiomahdollisuus sekä vertaisarviointi. Etenkin toimiva haku on tärkeä, koska jos ei löydä sellaista tietoa, joka olisikin tallennettu tietopankkiin, rapauttaa se käyttöaikomusta ja käyttöä tulevaisuudessa. Kaikki kartoitetut ominaisuudet löytyvät selitysten kanssa taulukosta 2.

Palvelun laatu huomioitiin ainoastaan vastuutahon nimeämisen kautta. Selkeä vastuutaho on kuitenkin tärkeässä roolissa tietopankkien ja myös koko tietämyksenhallinnan näkökulmasta. Sen kautta on mahdollista jatkokehittää organisaation tietämyksenhallintaa uusien hankkeiden kautta toivottuihin suuntiin helpommin.

Käyttöaikomus ja käyttö olivat laajalti esillä aineistossa. Paljon pohdittiin muun muassa miten välttää matala käyttöaste, vaikka tietopankissa olisikin sisältöä. Työn tuloksissa esitetään tähän kaksi isompaa varteenotettavaa ajatusta. Ensinnäkin tietopankin käyttö tulisi määritellä ja ottaa osaksi datatieteellistä prosessia, jotta toimintaa saataisiin ohjattua sen käyttöä kohti. Toisena havaintona oli tietopankin käytön mukana tulevan kulttuurillisen muutoksen huomioiminen. Tämä vaatii kuitenkin tarkempia jatkotutkimuksia konkreettisten ehdotuksien esittämiseksi. Työn perusteella voidaan kuitenkin sanoa, että tietopankkia ei tulisi ottaa käyttöön laajan lanseerauksen kautta ja vaan toivoa, että ihmiset lisäisivät sinne sisältöä.

Käyttäjätyytyväisyyden uskotaan kaikessa yksinkertaisuudessaan tulevan siitä, kun tietopankin käyttö koetaan omassa työssä hyödylliseksi. Tarkemmin sanottuna se vaatii laadukkaan ja oikeanlaisen sisällön sekä halutut ominaisuudet. Myös sen, että muut kokevat tallennetun tiedon hyödylliseksi arvioitiin vaikuttavan tyytyväisyyteen.

Tietopankista ja sen käytöstä saatavat nettohyödyt jaoteltiin tulosten perusteella organisaation hyötyihin ja yksilöiden hyötyihin. Organisaation näkökulmasta tietopankki takaa yhtenäiset työtavat ja asiakkaiden paremman palvelemisen. Yksilötasolla puolestaan tietopankki auttaa datatieteellisissä projekteissa etenemistä ja ongelmatapausten ratkomista sekä tukee oppimista.

Tuloksia tarkastellessa on jokaisen yksittäisen kokonaisuuden omien havaintojen lisäksi huomioitava niiden linkittyminen toisiin kokonaisuuksiin. Esimerkiksi sen lisäksi, että tietopankin käytön määritteleminen osaksi datatieteellistä prosessia vaikuttaa käyttöaikomukseen ja käyttöön, laadukkaan informaation löytyminen usealta eri sisältötasolta vaikuttaa siihen myös.

6. POHDINTA

Pohdinnassa tarkastellaan ensin työn tuloksia tutkimuskysymysten kautta tiiviisti aiempiin tutkimuksiin peilaten, jonka jälkeen käsitellään teoreettiset sekä käytännön kontribuutiot. Teoreettiset kontribuutiot keskittyvät tarkastelemaan korkeammalta tasolta miten tunnistettua tutkimusaukkoa käsitellään. Käytännön näkökulmasta puolestaan esitellään miten etenkin kodifiointistrategiaa seuraava organisaatio kykenee toteuttamaan tietämyksenhallintaa tietopankin kautta. Lisäksi tässä luvussa arvioidaan tutkimusta kokonaisuudessaan ja pohditaan jatkotutkimustarpeita.

6.1 Tulosten tarkastelu

Tässä osiossa vertaillaan työn tuloksia aiempaan tutkimukseen ja nostetaan esiin tämän työn mukana tuomia uusia näkökulmia ja rikastuksia aiempaan teoriaan. Tarkastelu on jaoteltu alatutkimuskysymysten perusteella helpommin käsiteltäviin kokonaisuuksiin. Ohessa pohditaan myös onnistuttiinko tutkimuskysymyksiin vastaamisessa ja millä tasolla. Lisäksi tämän alaluvun loppuun koostetaan taulukko tunnistettujen menestystekijöiden huomioimisesta.

Ensimmäinen alatutkimuskysymys *”mitä tietopankin tulisi sisältää, jotta se palvelee datatieteellisissä projekteissa toimimista?”* käsittelee tietopankin sisältöä. Tietopankin sisällölle ehdotetaan aineiston perusteella muovautuneet sisältötasot. Sisältötasot kuvaavat datatieteellisessä projektissa hyödynnettäviä resursseja eri abstraktiotasoilla, joita ovat projektitaso, konseptitaso ja käytäntötaso.

Sisältöön liittyen Haertelin et al. (2022) koostamat artefaktit ovat relevantteja. Tuloksista tietopankin konseptuaaliselta tasolta tunnistettiin sama Haertelin et al. (2022) esittämä infrastruktuurikuvaus ja projektitasolta puolestaan lopullinen projektiraportti. Muiden dokumenttien ja artefaktien puutuminen aineistosta selittyy yksinomaan tarkastelutasolla. Tutkimuksen tarkoituksena ei ollut syvällisesti paneutua tallennettavaan sisältöön, vaan esitellä yleiskatsaus siitä. Pienikin yhteneväisyys kertoo kuitenkin siitä, että tässä työssä kuvaillulla tietopankilla on oikea suunta sisältönsä puolesta. Lisäksi se tarjoaa tarkemman määrittelyn abstraktiotasoista, joita voidaan varmasti yhteensovittaa Haertelin et al. (2022) dokumenttien ja artefaktien kanssa. Toisaalta laaja kirjo erinäisiä dokumentteja, joita Haertel et al. (2022) mainitsee voi olla käytännössä liian raskasta toteuttaa, minkä vuoksi – ainakin tietopankin käytön alkuvaiheessa – tämän työn tiiviimpi lähestyminen sisältöön voi olla käytännössä toimivampi.

Myös työn tarjoamien sisältötasojen alimmalle eli käytännölliselle sisällölle kaavaillut sapluunat ovat sellaisia, joita kirjallisuudessa on aiemmin tästä aihepiiristä hieman käsitelty. Sapluunat ja niiden uudelleenkäyttäminen on Berinaton (2019) mukaan yksi edellytys onnistuneelle ja tehokkaalle datatieteelliselle projektille. Hänen näkemyksiensä perusteella tämä alin sisältötaso ja sen sisältämät funktionaaliset moduulit sekä sapluunat ovat relevantteja datatieteellisiä projekteja tukevalle tietopankille. Yhtä lailla tällä kyetään välttämään ongelmia, joita Ribièren ja Calabresen (2016) mukaan yksinomaan dokumenttikeskeisellä tietopankilla on. Tietopankin sisällön määrittelyyn tulee panostaa (Subramani et al., 2021; Veeravalli & Vijayalakshmi, 2021) ja näin myös tässä tutkimuksessa tehdään, joten ensimmäiseen alatutkimuskysymykseen onnistutaan vastaamaan.

Toisen alatutkimuskysymyksen kautta *”mitä ominaisuuksia tietopankista tulee löytyä?”* kartoitettiin teknisiä ominaisuuksia, joita tietopankki tarvitsee. Ominaisuuksia listattiin taulukkoon 2 kaksitoista (12) kappaletta, mutta tavoiteltujen ominaisuuksien lista tulee kuitenkin varmasti kasvamaan, kun ensimmäisiä versioita käytännön toteutuksesta ilmaantuu. Hakutoiminto on yksi selkeimmistä ominaisuuksista, joka on esillä sekä kirjallisuudessa (Chhim et al., 2017; Hetey et al., 2020; Subramani et al., 2021) että tuloksissa, ja myös versionhallinnasta on ollut mainintoja (Hetey et al., 2020). Kirjoittajan parhaan tietämyksen mukaan aiemmassa kirjallisuudessa ei ole kuitenkaan vastaavalla tarkastelutasolla esitelty datatieteellisiä projekteja tukevan tietopankin muita teknisiä ominaisuuksia, joten tässä työssä tehty selvitys tarjoaa uutuusarvoa tähän ja etenkin tarvittavan tietopankin käytännön toteutusta ajatellen.

Kolmas ja viimeinen alatutkimuskysymys *”mitkä tekijät ajavat tietopankin käyttöä?”* tarttui alhaisen käyttöasteen ongelmaan, joita historiallisilla tietopankeilla on ollut. Tulosten perusteella, kun tietopankin käyttö nähdään osana arkea, niin korreloi se korkeampaan käyttöasteeseen. Tämä tukee täysin Kankanhallin et al. (2011) tutkimuksen löydöstä, että tietopankkiin tallentaminen tulisi tapahtua osana normaalia työprosessia. Tutkimuksessa esitellään versio CRISP-DM prosessimallista, jossa on mukana tietopankin käyttämiseen liittyvät vaiheet. Kuten jo aiemmin alaluvussa 3.1 huomioitiin, organisaatiot ajautuvat kehittämään omia versioitaan datatieteellisistä prosesseista yksilöllisiin tarpeisiinsa (Saltz, 2015). Tällä ehdotuksella pyritään siihen, että organisaatio, joka toteuttaa paljon datatieteellisiä projekteja ottaisi mukautetun prosessimallin käyttöön ja siten toisi tietopankin käytön lähemmäs työntekoa. Tämä on osittain linjassa Gökalpin et al. (2022) koostaman parannellun prosessimallin kanssa. Osittain vain siksi, koska heillä käsittelyssä on rajatumpi analytiikkakeskeinen projekti, kun taas tässä työssä puhutaan korkeammalla tasolla datatieteellisistä projekteista. Idea on kuitenkin sama: prosessin ede-

tessä tietoa tulee tallentaa tietopankkiin seuraavien projektien helpottamiseksi. Prosessimallin näkökulmasta myös Haertelin et al. (2022) esittämät dokumentit voidaan uudessa *Tiedon jakaminen* -vaiheessa luontevasti tallentaa tietopankkiin. Samaan tapaan aiemmin tallennettuja dokumentteja voidaan etsiä *Tiedon hankkiminen -vaiheessa*.

Tuloksissa huomioidaan myös tietopankin käyttöönottoa organisaatiossa ja todetaan, että sen tulisi tapahtua hallitusti, koska muuten on todennäköisempää kohdata alhaisempi käyttöaste. Aivan kuten Veeravalli ja Vijayalakshmi (2021) painottavat. Varsinainen käyttöönotto voisi tapahtua alkuun esimerkiksi innovaatioiden diffuusiteorian mukaisesti innovaattorien ja varhaisten omaksujien kautta, koska näissä ryhmissä ovat ne henkilöt ketkä aidosti haluavat saada muutosta aikaan (Tidd, 2010). Käyttöönottoon liittyen tuloksissa esitetty fasilitaattori tietopankille ja koko tietämyksenhallinnalle on sellainen rooli, jota kirjallisuudessa suositaan tietämyksenhallinnan onnistumisen takaamiseksi (Ribière & Calabrese, 2016).

Lisäksi tietopankin kokeminen hyödylliseksi ajaa käyttöä. Tämä on hankala asia yleistää, sillä jokaisella yksilöllä on varmasti erilaisia tarpeita ja odotuksia tietopankin osalta. Todettakoon kuitenkin, että tulosten perusteella hyödylliseksi kokeminen tapahtuu sitä kautta, että sisältö on helposti saavutettavissa ja se vastaa datatieteellisissä projekteissa ilmeneviä tarpeita. Käyttöaikomukseen vaikuttaa positiivisesti tulosten perusteella myös tieto siitä, että muut käyttäjät ovat hyötyneet juuri sinun lisäämästäsi tiedosta, minkä myös Kankanhalli et al. (2005) havaitsivat tutkimuksessaan tietopankkien käyttämiseen vaikuttavista tekijöistä. Myös tietopankin käytön kautta tulevat nettohyödyt ovat linjassa aiempien tutkimuksien havaintoihin muun muassa työnteon nopeutumisesta ja oppimisen tukemisesta (Subramani et al., 2021).

Viimeisenä käsittelyssä on päätutkimuskysymys *”millainen tietopankki mahdollistaa datatieteellisten projektien toistettavuuden?”*. Onnistuneista vastauksista alatutkimuskysymyksiin on selkeästi koottavissa vastaus päätutkimuskysymykseen. Tietopankki, joka sisältää datatieteelliseen projektiin tarvittavia resursseja eri abstraktiotasoilla, on rakennettu käyttöä tukevista ominaisuuksista kuten hakutoiminnosta, ja on huomioitu datatieteellisessä prosessimallissa, mahdollistaa datatieteellisten projektien toistettavuuden. Kuvaillun tietopankin perusteella päätutkimuskysymykseen onnistutaan diplomityössä vastaamaan.

Toiset tutkimukset, joissa on lähdetty DeLonen & McLeanin (2003) viitekehyksen pohjalta analysoimaan tietopankkien onnistumista ovat tämän työn tulosten tarkastelun kannalta yksi keskeinen osa. Tällaisia ovat Wun ja Wangin (2006) sekä Filierin ja Willisonin (2016) tutkimukset, joiden onnistumisen tekijöistä suurinta osaa on käsitelty myös tämän

työn tuloksissa. Näitä huomioituja tekijöitä ovat muun muassa tietopankin sisällön kattavuus, laadukas sisältö, hyötyjen havainnoinnin tärkeys ja integraatiokyky. Huomioimatta jäi puolestaan tietopankin joustavuus. Vaikka tätä ei selkeästi mainittu, voidaan väittää, että työssä kartoitetut ominaisuudet edesauttavat sen saavuttamista. Esimerkiksi joustavuuteen vaikuttavat versionhallinta ja laatutasot. Näihin samankaltaisiin tutkimuksiin tämän työn tulokset ennen kaikkea syventävät tarkastelutasoa ja pilkkovat isompia kokonaisuuksia osiin.

Kattavan tulosten tarkastelun osalta on kiinnostavaa arvioida miten työn tulosten perusteella kuvailtu tietopankki kokonaisuudessaan huomioi menestystekijät, joita kirjallisuudessa on tietopankeille esitetty. Tunnistetut menestystekijät löytyvät alaluvusta 2.4. Menestystekijöiden huomioiminen ja tarkemmat keinot niihin ovat arvioituna alla olevassa taulukossa 3.

Taulukko 3. *Tietopankin menestystekijöiden huomioiminen työn tuloksissa*

Menestystekijä	**Huomioitu tuloksissa**	**Keinot**
Tieto on löydettävissä	X	- Hakutoiminto (järjestelmän laatu) - Kategorisointi (järjestelmän laatu) - Sisältötasot (informaation laatu) - Viittaukset/linkitykset *(järjestelmän laatu)*
Tietopankin sisältö on laadukasta	X	- Vertaisarviointi (järjestelmän laatu) - Kollaboraatio (järjestelmän laatu) - Tallennusvaiheen sapluuna *(järjestelmän laatu)*
Hyödynnetään semanttisia teknologioita		
Tuetaan uudenlaisia työtapoja	X	- Tietopankki osana datatieteellistä prosessia *(käyttöaikomus ja käyttö)*
Tietopankin sisältö on kattavaa	X	- Sisältötasot (informaation laatu)

		- Viittaukset/linkitykset *(järjestelmän laatu)* - Sisältöpyynnöt *(järjestelmän laatu)*
Tietopankin käyttöönotto hallitusti	X	- Käyttöönotto innovaattorien ja varhaisten omaksujien kautta *(käyttöaikomus ja käyttö)* - Fasilitaattori edistämässä *(palvelun laatu)*
Dokumenttikeskeisyyden välttäminen	X	- Käytännöllinen sisältötaso *(informaation laatu)*
Tietopankki on integroituva muihin järjestelmiin	X	- Huomioitava teknisen toteutuksen yhteydessä *(järjestelmän laatu)*
Tietopankki on joustava		
Hyötyjen tulee olla havaittavissa	X	- Koettujen hyötyjen viestiminen *(käyttäjätyytyväisyys)*

Teoriaosuudessa esille tuotuja menestystekijöitä tietopankin onnistumisen taustalla on tässä työssä saatu empiirisen osuuden kautta kattavasti huomioitua. Ainoat taulukon 3 menestystekijät, jotka eivät nousseet tuloksien kautta esille olivat semanttiset teknologiat ja tietopankin joustavuus. Puutteellisuuteen vaikutti vahvasti työn tarkastelun rajallisuus. Työssä ei esimerkiksi selvitetty konkreettisia menetelmiä ja teknologioita, kuten semanttisia teknologioita, joita tietopankin toteutuksen taustalla mahdollisesti tarvittaisiin. Joustavuutta tietopankin yhteydessä ei myöskään käsitelty suoranaisesti, mutta sen voidaan aiempien tutkimusten perusteella olettaa olevan tarpeen, kun tietopankin käyttöä sovitetaan käytännössä datatieteellisen projektin prosessimalliin.

Taulukosta 3 on havaittavissa, että työn tuloksissa tarjotut keinot menestystekijöiden huomioimiseksi ovat tasaisesti jakautuneet. Ei ole mitään yhtä selkeää ratkaisua onnistuneeseen tietopankkiin. Järjestelmän laatu -kokonaisuus on kuitenkin yksittäisenä isompana kokonaisuutena menestystekijöiden huomioimisen taustalla. Lisäksi useampi menestystekijä käsittelee tietopankin sisältöä, joten on myös luonnollista, että tuloksissa esitetyt sisältötasot ovat useampaan otteeseen esitetty menestystekijän huomioimisen keinona.

6.2 Teoreettiset kontribuutiot

Syvällisempi vertailu aiempaan teoriaan tehdään edellisessä luvussa ja tässä luvussa koostetaan siitä tiivis yhteenveto. Tämä työ seuraa Nakashan ja Bouhnikin (2021) näkemyksiä siitä, että tietämyksenhallintaan tulisi panostaa nykyään ja tulevaisuudessa uusien teknisten toteutusten ja teknologioiden avulla. Tässä työssä ei kuitenkaan selvitetä tai ehdoteta varsinaisia teknologioita tuloksissa esitettyjen ominaisuuksien saavuttamiseksi, vaan sen sijaan tarjotaan perusta sekä tarpeet uudenlaiselle datatieteellisiä projekteja tukevalle tietopankille.

Datatieteelliset projektit ovat ajoittain toistettavuuden kannalta hankalia (Martinez et al., 2021). Työssä esitellään ehdotuksia aiemmissa tutkimuksissa tunnistetun toistettavuushaasteen helpottamiseksi datatieteellisissä projekteissa. Selkeän tiedon uudelleenkäytön esilletuonti CRISP-DM prosessissa tiedon hankinnan ja tiedon jakamisen kautta tarjoaa tähän keinoja sekä potentiaalisiin jatkotutkimuksiin lähtökohtia. Tiedon uudelleenkäytön tehostamisen lisäksi tärkeimmät teoreettiset kontribuutiot työltä ovat tarkennukset sisällöstä ja teknisistä ominaisuuksista tietopankille, jota datatieteellisten projektien tueksi kirjallisuudessa kaivataan.

6.3 Käytännön kontribuutiot

Tietämyksenhallinta on laaja kokonaisuus, joka pitää sisällään muun muassa strategisia valintoja, tiedon jakamista ja kulttuurillisia muutoksia. Tässä työssä paneudutaan tiettyyn spesifiin – tietopankin hyödyntämisen – keinoon edistää organisaation tietämyksenhallintaa, minkä avulla luodaan myös eräänlainen kiintopiste tai lähtötilanne myöhemmin tulevalle ja laajemmalle tietämyksenhallinnalle. Tapaustutkimuksen perusteella tietämyksenhallinnalla on hyvät lähtökohdat case-organisaatiossa, koska sille on tunnistettu oikea tarve liiketoiminnan ääreltä ja haastateltavat vaikuttivat aidosti kiinnostuneilta asian jatkokehittämisestä.

Tämä tapaustutkimus tarjoaa kodifiointistrategiaa noudattavalle tai tavoittelevalle ja datatieteellisten projektien parissa toimivalle organisaatiolle suuntaviivat tietopankin toteuttamiselle. Tietopankin teknistä toteutusta voidaan aloittaa esimerkiksi tuloksissa esiteltyjen ominaisuuksien kautta. Tämä työ kertoo mitä tietopankilta odotetaan sen potentiaalisten käyttäjien mielestä, mihin on täten luotettava ja helppo tukeutua käytännön valintojen kanssa.

Tulosten kautta esitetään tietopankin sisällöstä konkreettisia ehdotuksia, jotka perustuvat tapaustutkimuksen kohteena olevan organisaation työntekijöiden mietteisiin. Käytännön näkökulmasta kohdeorganisaatio voi aloittaa tarkemman kartoittamisen eri sisältötasoille tallennettavista entiteeteistä ja samalla luoda niin sanotun minimisisällön. Sisällön selkeä määritteleminen on tärkeää, sillä juuri sen uudelleenkäyttö luo arvoa. Sen lisäksi työssä kuvaillaan kuinka organisaation tulisi huomioida tietopankki, jotta sen käyttö saadaan osaksi jokapäiväistä työntekoa. Edellä mainittu on erityisen tärkeä aspekti, sillä vaikka kuinka hienon teknisen toteutuksen toteuttaa, mutta se ei löydä paikkaa työnteossa, tietämyksenhallinnan aloite ei onnistu.

6.4 Tutkimuksen arviointi ja rajoitteet

Tutkimus on relevantti, sillä tutkittava aihe on ajankohtainen teoreettisen panoksensa kautta datatieteellisten projektien toistettavuusongelmaan ja myös käytännön panoksensa kautta tapaustutkimuksen kohdeyritykselle. Laadullisen tutkimuksen laatua voidaan arvioida tarkemmin muun muassa sen uskottavuuden, siirrettävyyden, luotettavuuden sekä vahvistettavuuden perusteella (Guba, 1981).

Uskottavuuden osalta pohditaan, että tutkitaanko tutkimuksessa sitä mitä on tarkoitus tutkia (Guba, 1981). Tässä tutkimuksessa tarkoituksena on kuvailla tietopankkia datatieteellisten projektien tukena ja näin myös tapahtuu, joten tämä kriteeri täyttyy. Tuloksia myös vertaillaan ja yhdistetään teoriaan, mikä osaltaan lisää uskottavuutta (Shenton, 2004). Siirrettävyys puolestaan käsittelee, että voidaanko tutkimuksen tuloksia soveltaa toisessa kontekstissa (Guba, 1981). Tietopankkien käyttö on selkeästi rajattu koskemaan datatieteellisisä projekteja, joten sen kontekstin ulkopuolella tämän tutkimuksen kaikkia tuloksia on mahdollisesti hankalampi soveltaa.

Seuraavana kriteerinä Guban (1981) nelijaossa on luotettavuus, jonka avulla otetaan kantaa tuloksien riippumattomuuteen. Tapaustutkimuksen teko yritykselle vie osansa täydellisen riippumattomuuden suhteen. Tarkoittaen sitä, että toiselle kohdeorganisaatiolle toteutettu vastaava tutkimus nostaisi todennäköisesti esiin joitakin eroja. Uskottavaa kuitenkin on, että kuka tahansa samalla haastattelurungolla (liite A) ja samojen haastateltavien kanssa saisi kerättyä samanlaisen aineiston. Johtopäätökset aineistosta saattaisivat toki hieman tässäkin painottua suuntaan tai toiseen, sillä etenkin pragmaattista tutkimusfilosofiaa noudattelevan tutkimuksen toteuttamisessa tutkijan omilla näkemyksillä on havaittu olevan vaikutusta.

Viimeisenä kriteerinä on vahvistettavuus, joka ottaa kantaa tulosten perustumiselle aineistoon (Guba, 1981). Tulosten vahvistettavuuden lähde on kuitenkin alkujaan jo haastattelurungon ja haastatteluhetken äärellä. Riskejä haastattelurungon kanssa oli muun muassa loppuun lisätyt herättelevät tarkennukset ja esimerkit muun muassa tietopankin sisällöstä. Näitä kuitenkin käytettiin pääosin vasta, jos keskustelua ei muuten syntynyt. Aineiston analysoinnin jälkeen kuitenkin huomataan, että tarkennukset ja lisäykset haastattelurungossa eivät juurikaan ole tuloksissa nähtävissä. Lisäksi haastatteluhetkessä objektiivisena pysyminen oli ajoittain haasteellista, koska kirjoittaja on myös käsittelyssä olevan tietopankin potentiaalinen käyttäjä. Pitäytyminen rungon kysymyksissä ja omien mielipiteiden minimointi piti tilanteen kuitenkin mahdollisimman objektiivisena.

Rajoitteita tutkimukselle asetti haastattelujen otannan koko. Vaikka se onnistuneesti sisälsi henkilöitä useista rooleista, kooltaan se olisi voinut olla isompi. Laajemmalla otannalla olisi voitu saavuttaa syvällisempää analyysiä ja mahdollisesti havainnoida tyypittelyn kautta mitä eri roolit ajattelevat tietopankista, sen sisällöstä ja käytöstä.

6.5 Jatkotutkimustarpeet

Jatkotutkimusaiheita ja -tarpeita nousi työn pohjalta useampi esiin. Tietopankin käyttöönotto, datatieteelliset projektit/prosessit ja tietopankin varsinainen toteuttaminen olivat selkeitä tällaisia kokonaisuuksia. Mainittuja kohteita käsitellään tarkemmin seuraavaksi.

Kuten tuloksissa mainittiin ja kirjallisuudessa on todettu, tietopankin saattaminen käyttöön tulee vaatimaan jonkun asteista kulttuurillista muutosta organisaatiossa. Jatkotutkimuksia ajatellen olisi tarpeen tutkia tarkemmin millaisesta kulttuurillisesta muutoksesta on kyse, miten se vaikuttaa sekä organisaatioon että yksilöihin ja miten se käytännössä saadaan aikaan. Kulttuurillinen muutos on suhteellisen laaja kokonaisuus, joten useampi tutkimus aiheesta on tarpeen.

Datatieteelliset prosessit ovat myös kiinnostava jatkotutkimuskohde. Esimerkiksi tarkemman selostuksen kokoaminen, miten tietopankkiin kytkeytyvät vaiheet datatieteellisessä prosessissa käytännössä toimivat, sopii yhdeksi jatkotutkimuskohteeksi. Lisäksi luonnollisena jatkumona on tutkimus siitä, mikäli tässä työssä kuvaillun tietopankin sisällyttäminen datatieteelliseen projektiin on merkittävää toistettavuuden parantamiselle. Toistettavuusongelmaan vastaamisen lisäksi asiaa on syytä tutkia selkeästi vain datatieteellisten projektien näkökulmasta. Voisi esimerkiksi pohtia vaikuttavatko lisävaiheet tietopankin käyttämisestä projektien tavoitteiden saavuttamiseen tai työmäärään.

Koska tässä tutkimuksessa selvitettiin vasta reunaehtoja tietynlaiselle tietopankille, tarvitaan jatkotutkimus selvittämään, kuinka tällainen tietopankki toteutettaisiin. Se toimisi

tavallaan tämän työn tulosten ja mallien testaamisena. Toteutuksen jälkeen on suotavaa myös tutkia vastaako toteutettu tietopankki odotuksia, onko tietopankille käyttöä ja tarvitseeko ehdotettuihin onnistumisen edellytyksiin muutoksia.

LÄHTEET

A.F. Ragab, M., & Arisha, A. (2013). Knowledge management and measurement: A critical review. Journal of Knowledge Management, 17(6), 873–901. https://doi.org/10.1108/JKM-12-2012-0381

Ahmed, B., Dannhauser, T., & Philip, N. (2018). A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects. CEEC, 11–14. https://doi.org/10.1109/CEEC.2018.8674234

Aviv, I., Hadar, I., & Levy, M. (2021). Knowledge management infrastructure framework for enhancing knowledge-intensive business processes. Sustainability (Basel, Switzerland), 13(20), 11387. https://doi.org/10.3390/su132011387

Awad, E. M., & Ghaziri, H. M. (2004). Knowledge management. Prentice Hall.

Baldé, M., Ferreira, A. I., & Maynard, T. (2018). SECI driven creativity: The role of team trust and intrinsic motivation. Journal of Knowledge Management, 22(8), 1688–1711. https://doi.org/10.1108/JKM-06-2017-0241

Bergeron, B. P. (2003). Essentials of knowledge management (1st edition). J. Wiley.

Berinato, S. (2019). Data Science and the Art of Persuasion. Harvard Business Review, 126.

Bock, G.-W., Mahmood, M., Sharma, S., & Kang, Y. J. (2010). The Impact of Information Overload and Contribution Overload on Continued Usage of Electronic Knowledge Repositories. Journal of Organizational Computing and Electronic Commerce, 20(3), 257–278. https://doi.org/10.1080/10919392.2010.494530

Bornstein, M., Li, J., & Casado, M. (2020). Emerging Architectures for Modern Data Infrastructure. Future. https://future.com/emerging-architectures-modern-data-infrastructure/ (Haettu 9.4.2023)

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cao, L. (2017). Data Science: A Comprehensive Overview. ACM Computing Surveys, 50(3), 1–42. https://doi.org/10.1145/3076253

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Chhim, P. P., Somers, T. M., & Chinnam, R. B. (2017). Knowledge reuse through electronic knowledge repositories: A multi theoretical study. Journal of Knowledge Management, 21(4), 741–764. https://doi.org/10.1108/JKM-03-2016-0126

Choo, C. W. (2016). The inquiring organization: How organizations acquire knowledge and seek information. Oxford University Press.

Dalkir, Kimiz. (2011). Knowledge management in theory and practice (2nd ed.). MIT Press.

Davenport, T. (2015). Whatever happened to knowledge management? Wall St. J.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90(10), 70–128.

Davenport, T. H., & Prusak, Laurence. (1998). Working knowledge: How organizations manage what they know. Harvard Business School Press.

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. Journal of Management Information Systems, 19(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73. https://doi.org/10.1145/2500499

Elkjaer, B., & Simpson, B. (2011). Pragmatism: A lived and living philosophy. What can it offer to contemporary organization theory? (Vol. 32, pp. 55–84). Emerald Group Publishing Limited. https://doi.org/10.1108/S0733-558X(2011)0000032005

Emmert-Streib, F., Moutari, S., & Dehmer, M. (2016). The Process of Analyzing Data is the Emergent Feature of Data Science. Frontiers in Genetics, 7. https://doi.org/10.3389/fgene.2016.00012

Eskola, Jari., & Suoranta, Juha. (1998). Johdatus laadulliseen tutkimukseen. Vastapaino.

Farquhar, J. Dawes. (2012). Case study research for business. SAGE.

Filieri, R., & Willison, R. (2016). Antecedents of Knowledge Sourcing and Reuse from a Knowledge Repository in the Virtual Product Prototyping: The Role of Knowledge and System Quality Dimensions. Knowledge and Process Management, 23(2), 147–160. https://doi.org/10.1002/kpm.1512

Goswami, A. K., & Agrawal, R. K. (2022). It's a knowledge centric world! Does ethical leadership promote knowledge sharing and knowledge creation? Psychological capital as mediator and shared goals as moderator. Journal of Knowledge Management, 27(3), 584–612. https://doi.org/10.1108/JKM-09-2021-0669

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. Educational Communication and Technology, 29(2). https://doi.org/10.1007/BF02766777

Gökalp, M. O., Gökalp, E., Kayabay, K., Gökalp, S., Koçyiğit, A., & Eren, P. E. (2022). A process assessment model for big data analytics. Computer Standards and Interfaces, 80, 103585. https://doi.org/10.1016/j.csi.2021.103585

Haertel, C., Pohl, M., Staegemann, D., & Turowski, K. (2022). Project Artifacts for the Data Science Lifecycle: A Comprehensive Overview. In Tsumoto S., Ohsawa Y., Chen L., Van den Poel D., Hu X., Motomura Y., Takagi T., Wu L., Xie Y., Abe A., & Raghavan V. (Eds.), Proc. - IEEE Int. Conf. Big Data, Big Data (pp. 2645–2654). Institute of Electrical and Electronics Engineers Inc.; Scopus. https://doi.org/10.1109/BigData55660.2022.10020291

Hansen, M. T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? Harvard Business Review, 77(2), 106–187.

Hetey, L., Neefs, E., Thomas, I., Zender, J., Vandaele, A.-C., Berkenbosch, S., Ristic, B., Bonnewijn, S., Delanoye, S., Leese, M., Mason, J., & Patel, M. (2020). Development of a knowledge management system for the NOMAD instrument onboard the ExoMars TGO spacecraft. Aircraft Engineering, 92(2), 81–92. https://doi.org/10.1108/AEAT-12-2018-0310

Hislop, Donald. (2013). Knowledge management in organizations: A critical introduction (3rd ed.). Oxford University Press.

Hotz, N. (2023). What is CRISP DM? Data Science Process Alliance. https://www.data-science-pm.com/crisp-dm-2/ (Haettu 9.4.2023)

Kankanhalli, A., Lee, O.-K. (Daniel), & Lim, K. H. (2011). Knowledge reuse through electronic repositories: A study in the context of customer service support. Information & Management, 48(2), 106–113. https://doi.org/10.1016/j.im.2011.02.002

Kankanhalli, A., Tan, B. C. Y., & Wei, K.-K. (2005). Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation. MIS Quarterly, 29(1), 113–143. https://doi.org/10.2307/25148670

Kelleher, J. D., Tierney, B., & Pietiläinen, K. (2021). Datatiede. Terra Cognita.

Kidwell, J., Linde, K., & Johnson, S. (2000). Applying Corporate Knowledge Management Practices in Higher Education. Educause Quarterly, 23.

Lindner, F., & Wald, A. (2011). Success factors of knowledge management in temporary organizations. International Journal of Project Management, 29(7), 877–888. https://doi.org/10.1016/j.ijproman.2010.09.003

Maier, R. (2007). Knowledge Management Systems: Information and Communication Technologies for Knowledge Management (3. Aufl., p. xiv). Springer-Verlag. https://doi.org/10.1007/978-3-540-71408-8

Martinez, I., Viles, E., & G. Olaizola, I. (2021). Data Science Methodologies: Current Challenges and Future Approaches. Big Data Research, 24. Scopus. https://doi.org/10.1016/j.bdr.2020.100183

Meredith, J. (1998). Building operations management theory through case and field research. Journal of Operations Management, 16(4), 441–454. https://doi.org/10.1016/S0272-6963(98)00023-0

Nagashima, H., & Kato, Y. (2019). APREP-DM: a Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM. INT CONF PERVAS COMP, 555–560.

Nakash, M., & Bouhnik, D. (2021). Knowledge management is not dead. It has changed its appearance. And it will continue to change. Knowledge and Process Management, 28(1), 29–39. https://doi.org/10.1002/kpm.1655

Nonaka, Ikujiro., & Takeuchi, Hirotaka. (1995). The knowledge-creating company: How Japanese companies create the dynamics of innovation. Oxford University Press.

Petter, S., DeLone, W., & McLean, E. R. (2013). Information Systems Success: The Quest for the Independent Variables. Journal of Management Information Systems, 29(4), 7–62. https://doi.org/10.2753/MIS0742-1222290401

Polanyi, M. (1966). The Tacit Dimension. Garden City, NY: Doubleday.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1(1), 51–59. https://doi.org/10.1089/big.2013.1508

Puusa, A., Juuti, P., & Aaltio, I. (2020). Laadullisen tutkimuksen näkökulmat ja menetelmät. Gaudeamus.

Raudeliuniene, J., Albats, E., & Kordab, M. (2021). Impact of information technologies and social networks on knowledge management processes in Middle Eastern audit and consulting companies. Journal of Knowledge Management, 25(4), 871–898. https://doi.org/10.1108/JKM-03-2020-0168

Ribière, V., & Calabrese, F. A. (2016). Why are companies still struggling to implement knowledge management? Answers from 34 experts in the field (pp. 13–34). https://doi.org/10.1016/B978-0-12-805187-0.00002-4

Rotondo, A., & Quilligan, F. (2020). Evolution Paths for Knowledge Discovery and Data Mining Process Models. SN Computer Science, 1(2). https://doi.org/10.1007/s42979-020-0117-6

Russell, K. E., La Londe, R., & Walters, F. (2016). Social knowledge: Organizational currencies in the new knowledge economy (pp. 141–150). https://doi.org/10.1016/B978-0-12-805187-0.00010-3

Sahibzada, U. F., Latif, K. F., Xu, Y., & Khalid, R. (2020). Catalyzing knowledge management processes towards knowledge worker satisfaction: Fuzzy-set qualitative comparative analysis. Journal of Knowledge Management, 24(10), 2373–2400. https://doi.org/10.1108/JKM-02-2020-0093

Saltz, J. S. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. BigData, 2066–2071. https://doi.org/10.1109/BigData.2015.7363988

Saltz, J. S., & Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. BigData, 2355–2361. https://doi.org/10.1109/BigData.2017.8258190

Saunders, M., Lewis, P., & Thornhill, A. (2019). Research Methods for Business Students. Pearson Education, Limited.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. Education for Information, 22(2), 63–75. https://doi.org/10.3233/EFI-2004-22201

So, J. C. F., & Bolloju, N. (2005). Explaining the intentions to share and reuse knowledge in the context of IT service operations. Journal of Knowledge Management, 9(6), 30–41. https://doi.org/10.1108/13673270510629945

Soto-Acosta, P., & Cegarra-Navarro, J.-G. (2016). New ICTs for Knowledge Management in Organizations. Journal of Knowledge Management, 20(3), 417–422. https://doi.org/10.1108/JKM-02-2016-0057

Subramani, M., Wagle, M., Ray, G., & Gupta, A. (2021). Capability development through just-in-time access to knowledge in document repositories: A longitudinal examination of technical problem solving. MIS Quarterly: Management Information Systems, 45(3), 1287–1308. Scopus. https://doi.org/10.25300/MISQ/2021/15635

Sugumaran, V. (2016). Semantic technologies for enhancing knowledge management systems (pp. 203–213). https://doi.org/10.1016/B978-0-12-805187-0.00014-0

Szulanski, G. (1996). Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm. Strategic Management Journal, 17(S2), 27–43. https://doi.org/10.1002/smj.4250171105

Tidd, J. (2010). Gaining momentum managing the diffusion of innovations. Imperial College Press.

Tuomi, J., & Sarajärvi, A. (2018). Laadullinen tutkimus ja sisällönanalyysi (Uudistettu laitos.). Kustannusosakeyhtiö Tammi.

Veeravalli, S., & Vijayalakshmi, V. (2021). Revisiting Knowledge Management System Use: Unravelling Interventions that Nurture Knowledge Seeking. International Journal of Knowledge Management, 18(1), 1–25. https://doi.org/10.4018/IJKM.291707

Voss, C., Tsikriktsis, N., & Frohlich, M. (2002). Case research in operations management. International Journal of Operations & Production Management, 22(2), 195–219. https://doi.org/10.1108/01443570210414329

Wu, J.-H., & Wang, Y.-M. (2006). Measuring KMS success: A respecification of the DeLone and McLean's model. Information & Management, 43(6), 728–739. https://doi.org/10.1016/j.im.2006.05.002

Yin, R. K. (2009). Case study research: Design and methods (4th ed.). SAGE Publications.

LIITE A: HAASTATTELURUNKO

Aloitus

- Esittäytyminen + tutkimuksen luottamuksellisuus, nimettömyys ja tarkoitus. Saako nauhoittaa?
- Käsitteiden tietämyksenhallinta ja tietopankki (engl. *knowledge repository / knowledge base*) selventäminen

Tietopankeista ja dataprojekteista yleisesti

- Oletko joutunut tekemään useassa projektissa samoja ”yksinkertaisia” asioita riippumatta liiketoimintaympäristöstä/kontekstista?
    a. *[Jos **on**, niin:]*
        i. Mitä nämä asiat ovat olleet?
        ii. (Koetko, että sitä aikaa olisi voinut käyttää paremmin? Oliko se ns. turhaa aikaa?)
        iii. Tapahtuuko tätä yhtä säännöllisesti jokaisessa vaiheessa projektia?
    b. *[Jos **ei**, niin:]* Oletko kuullut, että kollegasi ovat?
- Onko sinulla ollut tietopankki aiemmin käytössäsi?
    a. *[Jos **on**, niin:]* Mihin tarkoitukseen tietopankki oli? Millainen kokemus se oli?
    b. *[Jos **ei**, niin:]* Onko ollut sellaista tapausta, kun olisit toivonut, että käytössäsi olisi ollut tietopankki?

Tietopankki käytännössä

Pyri seuraavien kysymysten kohdalla pohtimaan asioita siitä näkökulmasta, että käyttäisit keskitettyä tietopankkia useiden eri projektien/asiakkaiden konteksteissa ja myös aktiivisesti täydentäisit sitä. (Huomioi minkä tasoinen tietopankki kyseessä --> isompi kuin projektitaso!)

- **Tietämyksenhallinnan prosessi**
    a. Mitä koet tarpeelliseksi, kun haluat…
        i. …tallentaa uutta tietoa tietopankkiin?

ii. …hyödyntää tietopankkiin tallennettua tietoa?

- **Informaation laatu**
    a. Mitä tietoa tulisi tallentaa tietopankkiin? *
    b. Missä muodoissa tietoa voisi olla?
- **Järjestelmän laatu**
    a. [Jos ei listattu tietämyksenhallinnan prosessi -kohdassa, niin:] Mitä ominaisuuksia tietopankista tulisi löytyä? ***
    b. Mistä teknologioista uskot olevan hyötyä tietopankissa? **
- **Palvelun laatu**
    a. Uskotko, että tietopankille löytyy aktiivisia käyttäjiä?
    b. Miten tietopankista ja etenkin sen sisällöstä saadaan luotettava?
- **Käyttäjätyytyväisyys / Käyttö**
    a. Mikä motivoisi sinua tallentamaan tietoa tietopankkiin?
    b. Millaisia ongelmia tai vaaranpaikkoja näet tietopankin käyttämisessä?
    c. Miten mahdollisimman moni saadaan käyttämään ja päivittämään tietopankkia?
        i. [Taustaoletus: palvelu pysyy hengissä vain, jos sitä käytetään ja päivitetään aktiivisesti usean henkilön toimesta]
    d. Koetko, että keskitetty tietopankki olisi hyödyllinen nykyisissä työtehtävissäsi?
        *i. [Jos **on**, niin:]*
            1. Kuvaile miten siitä olisi hyötyä nykyisissä työtehtävissäsi.

Muita

- Entä tuleeko mieleen jotain muuta, jota haluaisit vielä sanoa?

Teemoja, joihin voi tarttua

- *Tietojärjestelmiin liittyviä ongelmia:* pelkkä dokumenttikeskeinen lähestyminen ei toimi KM:n kanssa, keskitytään teknologioihin ihmisten sijasta, pettymys tietojärjestelmien kyvykkyyksiin, ihmisten kyvyt hyödyntää tarvittuja teknologioita ovat

puutteelliset, KM prosessissa keskitytään tiedon jakamisen mahdollistamiseen liikaa (muut myös tärkeitä)

Tarkennukset & Lisäykset

***** = arkkitehtuurikuvia, puhtaasti koodia, plain-text/markdown dokumentteja, IaC templaatteja, videoita, laskurit, ääniä…

** = versionhallinta (GitHub, ”medallion” arkkitehtuuri, tms.), luonnollisen kielen analyysi…

*** = API rajapinta, hakutoiminto, (data)katalogi, CRUD (create, read, update & delete), visualisointi, autom. luokittelu (esim. semantiikkaa)