



Vaasan yliopisto  
UNIVERSITY OF VAASA

Pasi Kytölä

***Yrityksen ulkoisen datan hyödyntäminen data-  
analytiikassa***

Tekniikan ja innovaatiojohtamisen yksikkö  
Tietojärjestelmätieteen  
pro gradu – tutkielma  
Digitaalisen liiketoiminnan kehittäminen

Vaasa 2020

---

**VAASAN YLIOPISTO****Tekniikan ja innovaatiojohtamisen yksikkö**

<b>Tekijä:</b>	Pasi Kytölä	
<b>Tutkielman nimi:</b>	Yrityksen ulkoisen datan hyödyntäminen data-analytiikassa	
<b>Tutkinto:</b>	Kauppätieteiden maisteri	
<b>Oppiaine:</b>	Digitaalisen liiketoiminnan kehittäminen	
<b>Työn ohjaaja:</b>	Juho-Pekka Mäkipää	
<b>Valmistumisvuosi:</b>	2020	<b>Sivumäärä:</b> 81

---

**TIIVISTELMÄ:**

Yritysten kilpailuetu on tiiviimmin sidottuna saatavilla olevaan dataan sekä yrityksen kykyyn tuottaa datamassoista liiketoiminnalle arvokasta informaatiota. Nykypäivän liiketoiminnassa saatavilla olevan datan tehokas hyödyntäminen ei liity pelkästään kilpailukykyyn vaan myös yrityksen selviytymiseen markkinoilla. Yhdistämällä ulkoista dataa ja omaa liiketoimintadataa yritykset ovat lukemattomien mahdollisuuksien äärellä.

Tässä tutkimuksessa keskitytään yritysten saatavilla olevaan ulkoiseen dataan, tavoitteena selvittää millä tavoin yritykset voivat saavuttaa liiketoiminnallista hyötyä tarjolla olevasta ulkoisesta datasta. Tutkimuksen tarkoitus on tuottaa tiedeyhteisölle ajantasainen kuva yritysten saatavilla olevasta avoimesta datasta ja maksullisesta datasta. Tutkimus rajataan koskemaan Suomen liiketoiminta-alueita. Tutkimus toteutetaan toimeksiantona suurelle suomalaiselle ohjelmistoalan yritykselle ja tutkimuksessa syntyvä erillinen tuotos luovutetaan yrityksen käyttöön.

Tutkimuksessa käytettiin ekstensiivistä tapaustutkimusta. Tutkimuksen tietoaineiston kerääminen toteutettiin käymällä läpi avoimen datan jakelukanavat, Suomen kansallisen tason portaali (Avoindata.fi) sekä alueelliset dataportaalit. Tietoaineistoa etsittiin myös datan avoimuuden edistäjien sekä yritysten ja organisaatioiden verkkosivuilta ja palveluista. Lopuksi tutkimuksen tuloksia tarkasteltiin case-yrityksen näkökulmasta.

Tutkimuksessa kerätystä aineistosta suurin osa koski julkishallinnon tarjoamaa avointa dataa. 125 suurimman datan tarjoajan joukosta ensimmäinen yksityisen alan yritys on sijalla 33. Tutkimus osoittaa kaupunkien aktivoitumisen datan avaamisessa. Suomen kuusi suurinta kaupunkia ovat 16 suurimman datan tarjoajien joukossa. Ulkoisesta datasta on 93 % ohjelmallisesti hyödynnettävissä sekä 54 % aineistosta on saatavilla ohjelmistorajapinnan kautta. Suurimmat datakategoriat olivat Alueet ja kaupungit, Liikenne sekä Väestö ja yhteiskunta.

Case-tutkimuksen perusteella tuotoksena syntynyt tietoaineisto on jatkohyödynnettävissä. Tietoaineistoa voidaan hyödyntää tapauskohtaisesti saatavilla olevan ulkoisen datan selvitykseen.

Datan avaamiseen liittyy useita moniulotteisia lainsäädännöllisiä kysymyksiä, joiden vuoksi yritykset eivät välttämättä ole valmiita siirtymään datan avaamiseen. Muita syitä datan avaamattomuuteen ovat taloudellisten kannustimien puuttuminen, pelko kilpailuedun menettämisestä, epäluottamus siitä, että dataa käytetään sopimusehtojen mukaisesti, epätasapaino yritysten neuvotteluvoimissa, ja pelko siitä, että kolmannet osapuolet väärinkäyttävät tietoja.

---

**AVAINSANAT:** Ulkoinen data, toissijainen data, avoin data, maksullinen data

## **Alkusanat**

Haluan kiittää työn toimeksiantajaa aiheesta sekä Teemu Hämäläistä ajatuksista ja työn sparrauksesta. Kiitos Juho-Pekka Mäkipäälle Vaasan yliopistosta työn ohjauksesta.

E erityiset kiitokset kotijoukoille; Ullalle, Voitolle ja Miisalle pitkäjänteisyydestä sekä kannustuksesta opiskeluissa.

Pasi Kytölä

Tampereella 8.10.2020.

## Sisällys

1	Johdanto	6
1.1	Tutkimuksen tavoitteet ja tutkimuskysymykset	8
1.2	Rajaukset ja rakenne	10
2	Tutkimuksen keskeiset käsitteet	11
2.1	Data-analytiikka	11
2.2	Data-analytiikan kypsyyssmalli	12
2.3	Big Data	15
2.4	Datan lähteet	17
2.5	Big Datan kypsyyssmalli	18
2.6	Datan integrointimallit	21
3	Ulkoinen data	23
3.1	Avoin data	24
3.2	Avoin data Suomessa	27
3.3	Avoimen datan jakelukanavat	29
3.4	Datan avoimuuden edistäjät	31
3.5	Maksullinen data	32
4	Ulkoisen datan tarjoajat Suomessa	35
4.1	Avoimen datan tuottajat	35
4.2	Maksullisen datan tuottajat	41
5	Tutkimusmenetelmä	47
5.1	Aineisto	48
6	Tulokset	51
6.1	Case-tutkimus	56
7	Diskussio	63
	Lähteet	68

## Kuvat

Kuva 1.	Kuva alueellisen rikollisuuden kehityksen analytiikassa käytettävistä ulkoisen datan lähteistä.	57
Kuva 2.	Kuva ympäristökatastrofeihin varautumisen analytiikassa käytettävistä ulkoisen datan lähteistä.	59
Kuva 3.	Kuva ihmismassojen hallinnan analytiikassa käytettävistä ulkoisen datan lähteistä.	61

## Kuviot

Kuvio 1.	Tutkimuksen toteutuksen vaiheistus.	9
Kuvio 2.	Data-analyysin vaiheet (mukaillen Runkler, 2016, s. 3).	12
Kuvio 3.	Gartnerin analytiikan kypsyysmalli (mukaillen Gartner, 2017).	14
Kuvio 4.	Big Datan määritelmä (mukaillen Russom, 2011, s. 6).	16
Kuvio 5.	Yritysten datalähteet (mukaillen Hartmann ja muut, 2016).	18
Kuvio 6.	TDWI-Big Data analytiikan kypsyysmallin vaiheet (mukaillen Halper & Stodder, 2014).	20
Kuvio 7.	Avoimen datan kerääminen Suomesta EU-tasolle (mukaillen Koski ja muut, 2017).	28
Kuvio 8.	Maksullisen datan vaihtoehdot (mukaillen Verhoeven, 2017).	33
Kuvio 9.	Aineisto kategorioittain.	52
Kuvio 10.	Suurimmat datan tarjoajat.	53
Kuvio 11.	Datan ohjelmallinen hyödynnettävyys.	53
Kuvio 12.	Käytettävissä oleva rajapinta.	54
Kuvio 13.	Aineiston tiedostotyytit.	55

## Taulukot

Taulukko 1.	Tietoaineistoissa käytetyt kategoriat (mukaillen Avoindata.fi, n.d.).	49
-------------	---	----

## 1 Johdanto

Data on meneillään olevan, neljänneksi teolliseksi vallankumoukseksi nimetyn, teknologivetoisen murroksen polttoainetta (Jantunen & Koskinen, 2019). Vuosien 2011 ja 2018 välillä markkina-arvoltaan suurempien yritysten lista on kohdannut muutoksen, jossa listaa hallitsevat öljy-yhtiöiden sijaan data-analytiikkaa hyödyntävät alustayhtiöt (Rajakannas, 2018a). Toisin kuin öljyn määrä hupenee käytettäessä, datasta tulee käytön myötä entistä hyödyllisempää ja käytetty data paljastaa usein jatkosovelluskohteita (Barr, 2018).

Dataa on tarjolla yhä enemmän kaupallisista ja ei-kaupallisista lähteistä. Yritysten kilpailuetu on tiiviimmin sidottuna tarjolla olevaan dataan sekä yrityksen kykyyn tuottaa datamassoista liiketoiminnalle arvokasta informaatiota. (Brownlow ja muut, 2015; Jantunen & Koskinen, 2019.) Yritykset, jotka epäonnistuvat tietopääoman käyttöönnotossa, ovat vaarassa menettää kriittisen kilpailuedun ja lopulta saavutetun markkinaosuuden. Nykypäivän liiketoiminnassa saatavilla olevan datan tehokas hyödyntäminen ei liity pelkästään kilpailukykyyn vaan myös yrityksen selviytymiseen markkinoilla. (Brownlow ja muut, 2015.)

Merkittävänä sysäyksenä saatavilla olevalle datalle toimii jo huippunsa yhteiskunnan ja talouden mullistavana tekijänä saavuttanut digitalisaation ylimainonta. Itkosen (2015) mukaan ylimainonnan huippu on saavutettu, kun valtion laitos alkaa kirjoittaa blogia trendikkästä aiheesta, mikä siis toteutuu Itkosen kirjoituksessa. Digitalisaatio tarkoittaa tiedon tallentamista, siirtämistä ja käsittelyä tietokoneiden ymmärtämässä muodossa (Itkonen 2015). Digitalisaatio toimii täten yhä enemmän sellaisen datan mahdollistajana, jota voidaan hyödyntää liiketoiminnan eri vaiheissa (Jantunen & Koskinen, 2019).

Yritysten saatavilla oleva data voidaan jakaa sisäiseen ja ulkoiseen dataan. Yritysten sisäinen data käsittää yrityksen hallussa olevan oman datan, joka voidaan ottaa yrityksen tietojärjestelmistä. Sisäinen data voi myös olla tiettyä tarkoitusta varten yrityksen itse luoma data. Ulkoinen data käsittää yrityksen ulkopuolelta hankitun datan. Ulkoinen data voidaan hankkia ostamalla tai data voi olla vapaasti saatavilla. (Hartmann ja muut, 2016.)

Yritysten avain menestykseen on näiden kahden datan yhdistely. Yhdistämällä ulkoista dataa ja omaa liiketoimintadataa ollaan lukemattomien mahdollisuuksien äärellä. (Var-teva & Suvanto, 2017.)

Euroopan komission mukaan digitaaliteknologiat ovat viime vuosina muuttaneet ta-loutta ja yhteiskuntaa, sekä vaikuttaneet kaikkiin toimialoihin ja kaikkien eurooppalais-ten jokapäiväiseen elämään. Komission mukaan muutoksen ytimessä toimii nyt ja tule-vaisuudessa yhä keskeisemmin data. (European Commission, 2020.) Datan merkittävä rooli esimerkiksi kriisinhallinnallisissa tehtävissä on tullut esiin myös tämän tutkimuksen aikana. Maailmalle levinneen koronaviruksen (COVID-19) etenemistä kuvaavaa avointa dataa Suomessa jakaa Terveysten ja hyvinvoinnin laitos (THL, 2020). Data on tukenut Suo-men hallituksen päätöksiä poikkeustilaan siirtymisestä sekä vastaavasti poikkeustilan ja rajoitusten päättämisestä.

Koronaviruksen vaikutuksia ihmisten liikkumiseen on pystytty seuramaan kännyköiden sijaintitietojen perusteella muodostettavalla datalla (Auvinen, 2020). Seurantaan käytet-tävä data on Telian Crowd Insights -palvelun maksullista dataa (Telia, 2018.)

Valtiovarainministerin ja keskustan puheenjohtajan Katri Kulmunin ministerintehtäväs-tään 5.6.2020 eroon johtanut selvitys alkoi puolestaan Hansel Oy:n tarjoamasta avoimen datan palvelusta. Tutkihankintoja.fi-palvelun avulla kansalaiset ja yritykset saavat tietoa valtion ja kuntien hankinnoista (Tutkihankintoja.fi, n.d). Palvelun avulla Suomen Kuva-lehden toimittajat selvittivät Kulmunin käyttäneen kahden ministeriön rahaa esiintymis-koulutuksiin myös valmistautuessaan keskustan puheenjohtajakisaan kesällä 2019.

## 1.1 Tutkimuksen tavoitteet ja tutkimuskysymykset

Tässä tutkimuksessa keskitytään yritysten saatavilla olevaan ulkoiseen dataan. Tutkimuksen tavoitteena on selvittää millä tavoin yritykset voivat saavuttaa liiketoiminnallista hyötyä tarjolla olevasta ulkoisesta datasta. Tutkimuksen tarkoitus on tuottaa tiedeyhteisölle ajantasainen kuva yritysten saatavilla olevasta ulkoisesta datasta sekä koota yhteen julkisen sektorin ja yksityisten yritysten tarjoama avoin data ja maksullinen data, mitä yritykset voivat käyttää liiketoiminnallisen hyödyn saavuttamiseen.

Suomessa avoimesta datasta löytyy paljon julkishallinnon tutkimuksia liittyen datan avaamiseen sekä avoimen datan liiketoimintapotentiaaliin. Avoimeen dataan liittyy erityisesti pohdintaa datan käyttömaksujen vaikutuksesta datan käyttömääriin. Pohdinta koskee sitä, mikä on avoimen datan tarjonnasta perittävä kohtuullinen korvaus ilman että perittävä korvaus olisi esteenä avoimen datan käytölle.

Julkishallinnon avatessa datalähteitä, keskustelu on siirtynyt yksityisten yritysten ja muiden organisaatioiden tarjoamaan julkiseen dataan. Tiedon avaamisen edistämiseksi ja keskustelun ylläpitämiseksi on perustettu useita yhdistyksiä, hankkeita ja tapahtumia. Yksityisten yritysten datan avaaminen on vielä edennyt hitaasti, eikä tarjolla olevasta datasta löydy avoimen datan kaltaista jakelukanavia, jotka kokoaisivat datalähteet yhteen.

Tämä tutkimus on tarpeellinen koska tutkimuksessa muodostuvaa ulkoisen datan kokoaavaa listaa ei Suomen liiketoiminta-alueelta ole aiemmin toteutettu. Vaikka avattuja tietolähteitä on käsitelty erityisesti julkishallinnon näkökulmasta, yksityisten yritysten tarjoamaa avointa datasta ei ole koottu yhteen. Koottu tietoaineisto antaa ulkoisen datan hyödyntäjille konkreettista tietoa tarjolla olevasta tietoaineistosta sekä tietoaineiston hyödyntämismahdollisuuksista käyttökohteissa. Konkreettista tietoa ovat esimerkiksi vaihtoehdot käytettävissä olevista rajapinnoista ja tiedostomuodoista.

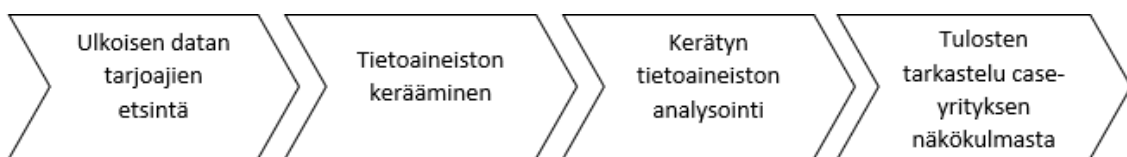
Tutkimuksen toimeksiantajana toimii muun muassa analytiikkapalveluita tarjoava suuri suomalainen ohjelmistoyritys. Toimeksiantajayritys tulee hyödyntämään tietoaineistoa



uusien analytiikkaliiketoimintamahdollisuuksien ideoinnissa sekä tulee käyttämään aineistoa asiakasprojektiensa tukena. Tietoaineiston perusteella toimeksiantajaryitys pystyy arvioimaan tarjolla olevaa ulkoista dataa osana asiakkaalle toimitettavaa analytiikkaratkaisua. Tutkimuksessa kerättävä tietoaineisto toimitetaan toimeksiantajaryitykselle koneellisesti käytettävässä muodossa.

Toimeksiannon taustalla on ajatus, että yleisellä tasolla ei ole vielä ymmärretty yrityksen ulkopuolisen datan tuomia mahdollisuuksia liiketoiminnan analytiikan tukena. Datamassojen avautuessa avautuu koko ajan uusia ja erilaisia tapoja hyödyntää ulkoista dataa.

Tutkimuksessa analysoidun tietoaineiston kerääminen toteutetaan käymällä läpi ulkoisen datan jakelukanavat Suomessa. Aineiston keräämisen ajankohtana oli maaliskuun ja kesäkuun 2020 välinen aika. Analysoidun aineiston tuloksia tarkastellaan case-yrityksen näkökulmasta. Tutkimuksen toteutus vaiheistetaan kuvion 1 mukaisesti.



**Kuvio 1.** Tutkimuksen toteutuksen vaiheistus.

Tutkimuksen tavoitetta lähestytään seuraavilla tutkimuskysymyksillä:

- Mitä avointa dataa ja maksullista dataa on tarjolla?
- Onko tarjolla oleva data ohjelmallisesti hyödynnettävissä?
- Mikä on yritysten esteenä tarjota avointa dataa ja maksullista dataa?

## 1.2 Rajaukset ja rakenne

Tutkimus rajataan koskemaan Suomen liiketoiminta-aluetta. Selvitettävä avoin data rajataan koskemaan Suomen lisäksi Euroopan avoimen datan portaalia. Maksullisen datan lähteet rajataan koskemaan Suomea.

Tutkimuksen teoriaosuudessa (luku 2) luodaan yhtenäinen ja tiivis katsaus datan käsitteeseen, datan lähteisiin, datan integraatiomalleihin sekä data-analytiikkaan. Kappaleen tarkoitus on määrittää työn aihealueeseen olennaisesti liittyvät peruskäsitteet.

Luvussa kolme kuvataan ulkoisen datan käsite sekä tehdään katsaus Suomen datan avoimuuden kehitykseen. Luvussa kuvataan lisäksi avoimen datan jakelukanavat ja maksullisen datan liiketoimintamallit.

Neljännessä luvussa kootaan yhteen avoimen datan ja maksullisen datan tietoaaineistot. Luvussa kuvataan avoimen datan ja maksullisen datan tarjoajat sekä tarjolla olevat tietoaaineistot. Viidennessä luvussa esitetään käytetty tutkimusmenetelmä sekä kerättävä aineisto.

Luvussa kuusi esitetään tutkimuksen tulokset, vastataan asetettuihin tutkimuskysymyksiin ja arvioidaan tutkimuksen tuloksia case-yrityksen näkökulmasta. Luvussa seitsemän esitetään tutkimuksen diskussio. Diskussiossa tiivistetään työn tärkeimmät havainnot ja analysoidaan tutkimuksen tulokset. Luvussa tehdään myös ehdotukset jatkotutkimuksista ja arvioitaan tutkimuksen reliabiliteettia sekä validiutta.

## 2 Tutkimuksen keskeiset käsitteet

Tässä luvussa käsitellään ulkoiseen dataan liittyvää teoreettista taustaa ja datan hyödyntämiseen liittyvää peruskäsitteistöä. Luvussa kuvataan datan lähteet sekä datan tarjoamiseen liittyvät liiketoimintamallit. Ulkoisen datan osana määritellään Big Data, memettä kuitenkin tarkemmin Big Datan rakenteen mukaisiin kategorioihin. Datan käsittelyyn liittyen määritellään Big Datan maturiteetti, joka kertoo yrityksen valmiutta käyttää hyväksi saatavilla olevaa suurta datamassaa.

Luku aloitetaan esittelemällä data-analytiikan konsepti, joka liittyy olennaisesti tutkimuksen tarkoitukseen selvittää yritysten liiketoiminnan tukemiseksi saatavilla oleva ulkoinen data. Saatavilla olevan datan lisäksi yrityksellä täytyy olla kyky tuottaa datasta liiketoiminnallisesti arvokasta informaatiota. Vastaavalla tavalla kuin öljy, data tarvitsee jalostamista ennen todellisen arvon hyödyntämistä (Barr, 2018).

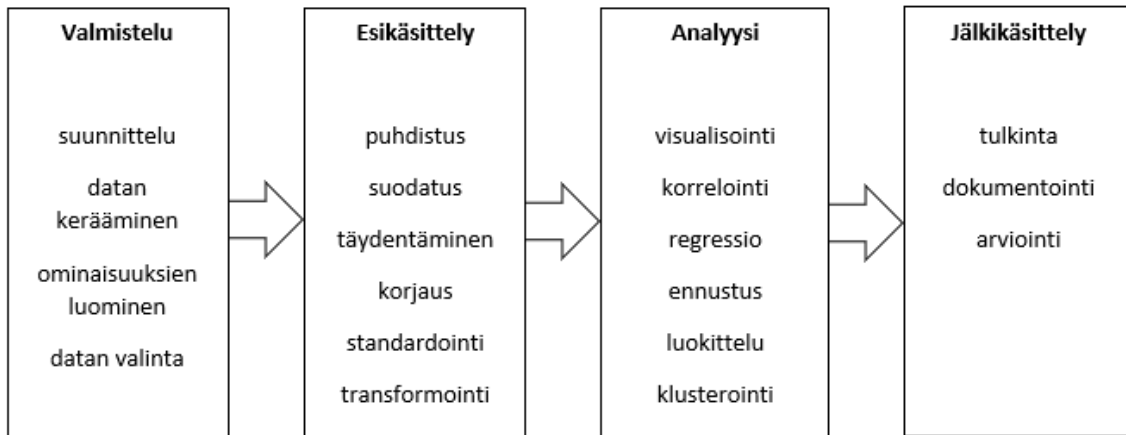
### 2.1 Data-analytiikka

Data-analytiikka on ollut käsitteenä olemassa 1960-luvulta asti, jolloin tilastotieteilijä John Tukey tulkitse data-analytiikan tarkoittavan tilastotieteellistä menetelmää datan keräämiseen, prosessointiin ja tulkintaan (Tukey, 1962, s. 2). Terminä data-analytiikka tuli suosituksi 2000-luvun alussa (Runkler, 2016, s. 2).

Data-analytiikan määritelmänä voidaan pitää suurten datamäärien hyödyntämistä päätöksenteon tukena. Suurten datamäärien hyödyntämisessä on oleellista soveltaa tietojärjestelmiä. Data-analytiikka on monitieteinen ala, joka on omaksunut näkökulmia tilastotieteestä, koneoppimisesta, hahmotunnistuksesta, systeemiteoriasta, operaatioanalyysistä sekä tekoälystä. (Runkler, 2016, s. 2.)

Data-analyysi on data-analytiikan jalkauttamista käytäntöön. Data-analyysi jaetaan tyypillisesti useaan vaiheeseen, joita ovat datan valinta, datan puhdistus, data visualisointi

ja datan analysointi. Analysoitu data tulkitaan ja arvioidaan. (Runkler, 2016 s. 2.) Runkler (2016, s. 2) ryhmittelee vaihteet valmisteluun, esikäsitteilyyn, analyysiin sekä jälkikäsitteilyyn (ks. kuvio 2).



**Kuvio 2.** Data-analyysin vaiheet (mukailen Runkler, 2016, s. 3).

Tässä tutkimuksessa keskitytään data-analyysin ensimmäiseen vaiheeseen, joka sisältää saatavilla olevan datan selvityksen, käytettävän datan valinnan sekä datan keräämisen.

## 2.2 Data-analytiikan kypsyysmalli

Kypsyys voidaan yksinkertaisimmillaan ymmärtää tilana, jossa kaikki on täysin valmista. Kypsyys taso viittaa aikajärjestyksessä toisiaan seuraaviin muutoksiin, joilla on ollut vaikutusta nykytilaan (Leem ja muut, 2008, s. 1201.) Kypsyysmalli koostuu puolestaan kypsyystasojen sarjasta. Kypsyysmalli edustaa ennakoitua, toivottua tai tyypillistä kehityspolkua ja on muotoiltu erillisiksi vaiheiksi. (Becker ja muut, 2009, s. 213.)

Kypsyysmalli kuvaa organisaation tai organisaation osan tilaa tai kehitysvaihetta. Kypsyysmallin avulla voidaan arvioida ja mitata organisaation prosessien kypsyystä, ja arvioinnin tuloksena organisaatio pyrkii parantamaan toimintatapaansa. Kypsyysmalli on

viitekehysten tyyppi, joka tukee organisaation kehittymistä. (Aho, 2011, s. 64.) Kypsyysmallista käytetään usein englanninkielistä vastinetta, *maturity* (maturity).

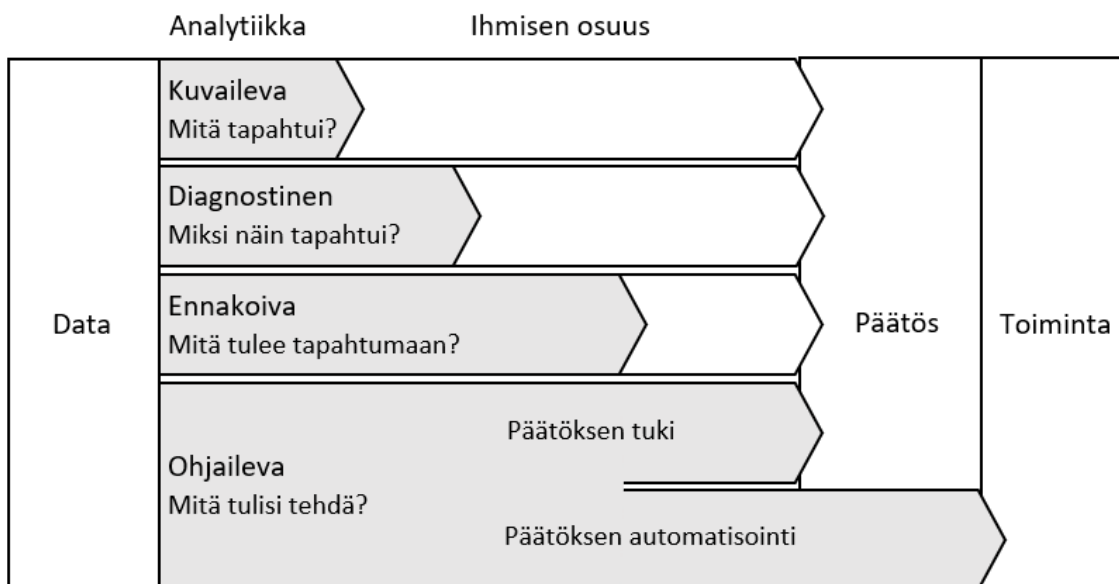
Kypsyysmallin konsepti kehitettiin 1970-luvulla. Konseptin alkuaajoista lähtien eri tarkoituksiin kehitettyjä kypsyysmalleja on tehty lukematon määrä (Mettler ja muut, 2010.) Mettler ja muut (2010) kävivät tutkimuksessaan läpi 117 tietojärjestelmätieteen alalle kehitettyä kypsyysmallia.

Ahon (2011, s. 64) mukaan kypsyysmalleja kehitettiin alun perin informaatioteknologian käyttöönotolle ja käytölle, ohjelmistoprosesseille sekä käytettävyyden huomioonottamiselle. Nykypäivän kypsyysmallit pohjautuvat Carnegie-Mellonin yliopiston ohjelmistokehitysinstituutissa 1980-luvun lopulla kehitettyyn prosessien kypsyysmalliin, *Capability Maturity Model (CMM)*. CMM-mallin kehitysasteet tarjoavat organisaatiolle suositellun järjestyksen lähestyttäessä vaiheittaista prosessin parannusta, tavoitteena muuttaa tiettyä tarkoitusta varten luotuja kehittymättömät prosessit kypsiin, laadukkaisiin ja tehokkaisiin prosesseihin. (CMMI, 2002, s. 1; Aho, 2011, s. 64, 67.)

Data-analytiikan kypsyysmalli ohjaa organisaatioita analytiikan hyödyntämisessä. Kypsyysmalli tarjoaa ymmärryksen organisaation analytiikan hyödyntämisen nykytilasta ja menneisyydestä sekä kuinka hyödyntämisessä tulisi edetä. Kypsyysmalli myös auttaa organisaatioita ymmärtämään analytiikan hyödyntämisen parhaita käytäntöjä muilta organisaatioilta. Kypsyysmalli tarjoaa organisaatiolle metodologian analytiikan hyödyntämisen mittaamiseen ja seuraamiseen sekä askeleet, joilla organisaatio saavuttaa kypsyysmallin seuraavan tason (Halper & Stodder, 2014).

Gartner (2017) kuvaa neljä analytiikan tasoa, jotka kuvastavat organisaation kypsyyttä käyttää hyväksi datasta saatavaa tietoa ja tehdä tiedon perusteella päätöksiä sekä toimia (ks. kuvio 3). Korkean kypsyystason saavuttaneilla organisaatioilla ihmisen osuus datan perusteella tehtäviin päätöksiin on pienempi kuin alhaisen kypsyystason organisaatioilla.

Analytiikan avulla päätöksenteko sekä päätöstä vastaava reagointi voidaan automatisoida.



**Kuvio 3.** Gartnerin analytiikan kypsyyssmalli (mukaillen Gartner, 2017).

Kuvaileva analytiikka vastaa kysymykseen ”Mitä tapahtui?”. Kuvaileva analytiikka vastaa perinteistä raportointia keräämällä yhteenvetoja tärkeimmistä suorituskykymittareista. Päätöksentekijät käyttävät raporttien lisäksi omaa arviointia ja kokemusta päätöksenteon tukena ja suorittavat päätösten perusteella tarvittavia toimenpiteitä. (Gartner, 2017.)

Diagnostinen analytiikka vastaa kysymykseen ”Miksi näin tapahtui?” ja on kuvailevaa analytiikkaan yksityiskohtaisempi, interaktiivinen analytiikka. Diagnostisessa analytiikassa käytetään sovelluksia, kuten visualisointi- ja tiedonkeruutyökaluja ja päätöksenteossa luotetaan enemmän analytiikkaan ihmisen osuuden pienentyessä. (Gartner, 2017.)

Ennakoiva analytiikka vastaa kysymykseen ”Mitä tulee tapahtumaan?” tunnistamalla datasta malleja ja arvioimalla todennäköisiä tuloksia tilastollisia menetelmiä sekä

koneoppimista. Ennakoiva analytiikka tarjoaa päätöksille paremman tuen tulevaisuuteen suuntaavilla näkemyksillä. (Gartner, 2017.)

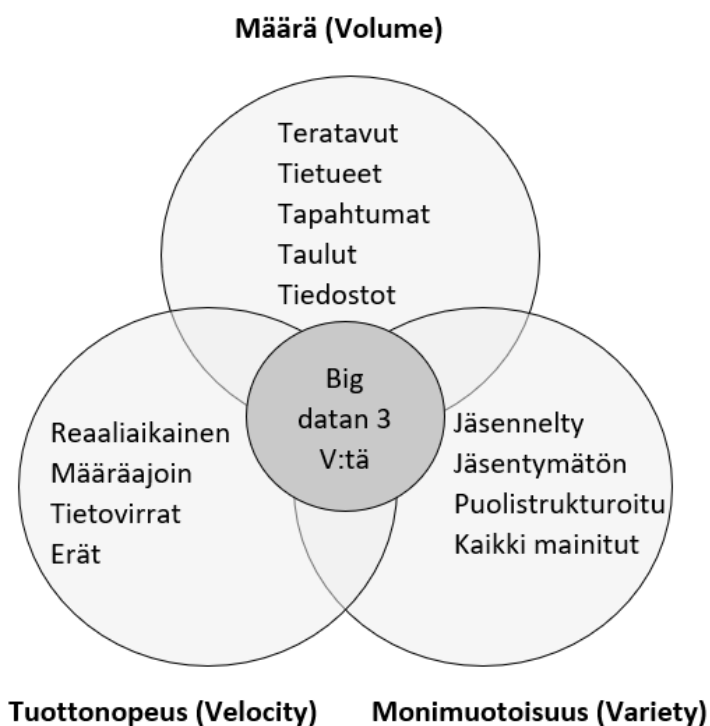
Ohjaileva analytiikka vastaa kysymykseen ”Mitä tulisi tehdä?” laskemalla päätöksentekovaihtoehtoihin liittyvät oletetut lopputulokset. Analytiikka tuottaa parhaimman tuloksen, kun päätöksen tavoitteet, rajoitukset ja perusteet on rajattu. Ohjailevaa analytiikkaa voidaan käyttää päätöksenteon tukena ja päätöksenteon automatisointiin. Päätöksen tukena analytiikka palauttaa päätösehdotuksen päätöksentekijälle, joka hyväksyy ehdotuksen tai muuttaa ehdotusta. Päätöksenteon automaatiassa analytiikan tuottama päätös toteutetaan järjestelmässä automaattisesti. (Gartner, 2017.)

### 2.3 Big Data

Toisin kuin polttoaineen pohjana oleva raakaöljy, analytiikan pohjana oleva Big Data sisältää paljon vaihtelua. Big Data voi sisältää sanoja, kuvia, ääntä, ideoita, faktoja, mittauksia, tilastotietoa tai jotain muuta koneellisesti informaatioksi muunnettavaa aineistoa. (Marr, 2018.)

Comuzzin ja Patelin (2016) mukaan Big Data on terminä suhteellisen uusi ja kuvaa sekä jäsenneilyn että jäsentämättömän datan eksponentiaalista kasvua ja saatavuutta. Datan katsotaan olevan Big Dataa, kun dataa ei voida käsitellä tämän hetken laajalle levinneellä tekniikalla, kuten relaatiotietokannat tai laskentataulukko-sovellukset (Comuzzi & Patel, 2016).

Big Datan määritelmäksi on vakiintunut Gartnerin kolmen v:n määritelmä (Ward & Barker, 2013). Gartnerin (n.d.) määrittelyn mukaan Big Data on suuren tietomäärän (volume), suuren tuottonopeuden (velocity) ja suuren monimuotoisuuden (variety) sisältäviä tietovirtoja (ks. kuvio 4). Tietovirtojen käsittely vaatii kustannustehokkaita ja innovatiivisia prosessointikeinoja paremman ymmärryksen tuottamiseksi päätöksenteon tueksi (Gartner, n.d.).



**Kuvio 4.** Big Datan määritelmä (mukaiillen Russom, 2011, s. 6).

Kolmen v:n lisäksi Big Datan määritelmään on lisätty muita ominaisuuksia. IBM lisäsi määritelmään neljännen v:n kuvaamaan datan totuudenmukaisuutta lähinnä datan epäluotettavuuden näkökulmasta (veracity) (Miele & Shockley, 2013). SAS lisäsi ominaisuuksien listaan datan vaihtelevuuden (variability) ja monimutkaisuuden (complexity). Vaihtelevuus kuvaa sitä, että datan tuottonopeus ei ole vakio ja sisältää jaksottaista vaihtelua. Monimutkaisuus kuvaa sitä, että data tuotetaan lukemattomista eri lähteistä. Oracle lisäsi Big Datan määritelmään datan arvon (value). (Gandomi & Haider, 2015.)

Datalla on usein suhteellisen alhainen arvo verrattuna datan määrään. Datan korkea arvo saavutetaan analysoimalla suuria tietomääriä. (Gandomi & Haider, 2015.) Data-analytiikan näkökulmasta Sun ja Huo (2019) listaavat Big Datan määritelmään vielä älykkyyden (big intelligence), analytiikan (big analytics), infrastruktuurin (big infrastructure), palvelun (big service) ja markkinat (big market).



## 2.4 Datan lähteet

Datalla tarkoitetaan digitaalisesti tallennettua, merkeistä ja symboleista koostuvaa ko-  
neellisesti luettavissa olevaa aineistoa, joka voi muodostaa dokumentteja, tilastoja, tie-  
tokantoja, kuvia, videotallenteita, audiotallenteita, karttoja ja 3D-malleja. Data voidaan  
ymmärtää raaka-aineena, jota jalostamalla syntyy merkityksellistä informaatiota. (Avoin-  
data.fi, n.d.)

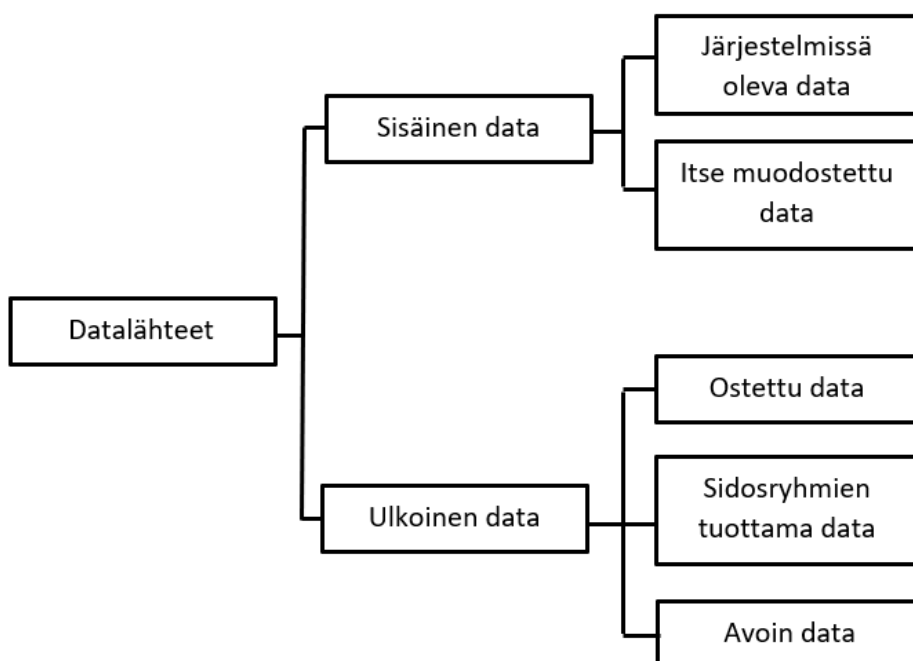
Dataa on saatavilla useista eri lähteistä. George ja muut (2014) sekä Hartmann ja muut  
(2016) jaottelevat data viiteen eri lähdetyyppiin. Hartmann ja muut (2016) tarkentavat  
datan lähteiksi järjestelmien tuottaman operatiivisen datan, yritysten omistaman mutta  
käyttämättömän, pimeän datan (dark data), kolmansilta osapuolilta hankitun kaupallisen  
datan, sosiaalisen datan sekä julkisen datan. George ja muut (2014) puolestaan jaottelevat  
lähteiksi julkisen datan, yksityisen datan, ylijäämädatan, yhteisödatan sekä itse kvan-  
tifioidun datan.

Ylijäämädata on ympäröivää passiivisesti kerääntyvää dataa, jolla ilman muuhun dataan  
tehtyä yhdistelyä on hyvin vähän tai ei ollenkaan arvoa. Ylijäämädataa ovat esimerkiksi  
kuluttajien tekemät ostokset sekä internethaut. Yhteisödata kuvaa sosiaalisia suuntauk-  
sia. Yhteisödataa ovat esimerkiksi tuotteiden kuluttaja-arvostelut sekä äänestyspainik-  
keilla kerättävä data. Itse kvantifioitu data on kuluttajien tuottama data. Itse kvantifioitua  
dataa on esimerkiksi aktiivisuusmittareilla kerättävä data, joka ladataan applikaatioihin.  
(George ja muut, 2014.) Pimeän datan, eli käyttämättömäksi jäävän datan määrä kerä-  
tystä datasta saattaa olla arvioiden mukaan jopa 90 prosenttia (Barker, 2016).

Hartmann ja muut (2016) tunnistavat saatavilla olevasta datasta kahdeksan tietolähdettä,  
jotka voidaan jakaa yritysten sisäisiksi ja ulkoisiksi tietolähteiksi (ks. kuvio 5). Sisäisiin tie-  
tolähteisiin kuuluvat yrityksen tietojärjestelmien sisältämä data sekä yrityksen tiettyä  
käyttötarkoitusta itse muodostama data. Tiettyä käyttötarkoitusta varten tuotettu data  
voi olla tuotettu seurantajärjestelmillä, kuten sensoreilla tai verkkonavigointilaitteilla.

Data voi olla tuotettu myös joukkoistamalla, jossa data kerätään laajalta osallistujajoukolta verkossa tai muilla yhteistyömenetelmillä.

Ulkoisen datan tietolähteitä ovat ulkopuolisilta tarjoajilta ostettu data, asiakkaiden tai liiketoimintakumppaneiden tuottama data, joka ei ole yleisesti saatavilla sekä vapaasti saatavilla oleva avoin data (Hartmann ja muut, 2016). CITO Research (2012) lisää listaan vielä organisaation omistaman mutta toisen osapuolen hallinnoiman datan (second-party data). Tällaista dataa ovat esimerkiksi sähköpostijärjestelmien tai pilvipalveluiden sisältämä data.



**Kuvio 5.** Yritysten datalähteet (mukaillen Hartmann ja muut, 2016).

## 2.5 Big Datan kypsyyssmalli

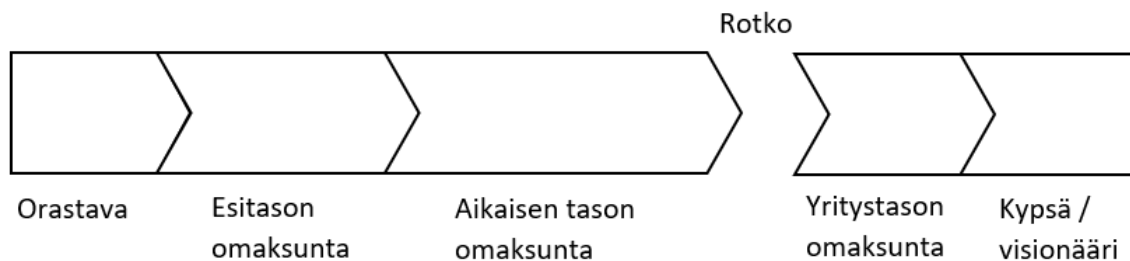
Gandomi & Haider (2015) mainitsevat, että tyhjiössä oleva Big Data on hyödytöntä. Tällä he tarkoittavat, että Big Datan potentiaali paljastuu vasta kun dataa hyödynnetään päätöksenteossa. Mahdollistaakseen dataohjautuvan päätöksenteon, organisaatiot

tarvitsevat tehokkaita prosesseja, joilla monimuotoinen Big Data muunnetaan merkitykselliseksi tiedoksi (Gandomi & Haider, 2015). Halperin & Krishnanin (2013) mukaan Big Datan kypsyyssmalli voidaankin määritellä organisaation evoluutioksi integroida, hallita ja hyödyntää olennaisia sisäisiä ja ulkoisia datalähteitä.

Suurten tietomäärien saatavuus tarjoaa organisaatioille useita mahdollisuuksia. Ensimmäinen vaikutus Big Datalla on organisaation päätöksentekoon. Organisaation kyky käsitellä nopeasti suuria tietomääriä antaa organisaatioille mahdollisuuden tehdä paremmin perusteltuja päätöksiä lyhyemmässä ajassa kilpailijoihin verrattuna. (Comuzzi & Patel, 2016.) Big Datan vallankumous ei johdu niinkään tarjolla olevan datan määrän räjähdysmäisestä kasvusta, vaan nykyisten tietojärjestelmien kyvystä käsitellä suuria datamääriä ja luoda datasta merkityksellistä informaatiota (Akerkar & Sajja, 2016, s. 31). Vielä 2000-luvun alussa Big Data muodosti tietojärjestelmille merkittävän ongelman, kun laitteiden tiedontallennus- ja prosessointitekniikat eivät pystyneet käsittelemään tehokkaasti valtavaa datamäärää (Russom, 2011, s. 4).

Organisaatiot tarvitsevat yhä merkittävän määrän ohjausta Big Datan tuottaman arvon hyödyntämiseen. Organisaatioiden tarvitsema ohjaus sisältyy tyypillisesti Big Datan kypsyyssmalliin, jota käytetään arvioimaan organisaatioiden nykyistä tilannetta ja määrittämään seuraavat vaiheet organisaatioiden aseman parantamiseksi lähitulevaisuudessa. (Comuzzi & Patel, 2016.)

Big Data -analytiikka on tekniikka Big Datan keräämiseen, datasta löytyvän tiedon tunnistamiseen, tiedon visualisointiin sekä tiedon tarjoamiseen päätöksenteon tueksi (Sun & Huo, 2019). The Data Warehouse Institute (TDWI) tarjoaa Big Data-analytiikan kypsyyssmallin, jossa organisaation Big Datan hyödyntämisasteet on jaettu viiteen tasoon (Halper & Stodder, 2014). Tasot esitetään kuviossa 6.



**Kuvio 6.** TDWI-Big Data analytiikan kypsyysmallin vaiheet (mukaillen Halper & Stodder, 2014).

Orastavalla tasolla (nascent) organisaatiolla ei ole tieto-ohjautuvaa kulttuuria ja päätökset tehdään enemmän vaiston kuin faktojen pohjalta (Halper & Stodder, 2014). Organisaatiossa saattaa olla yksittäisiä henkilöitä, jotka ovat kiinnostuneita Big Datan mahdollisesta potentiaalista. Organisaatio saattaa lisäksi olla tutustunut analytiikan konseptiin ja yrityksellä voi olla esimerkiksi tietovarasto, mutta varsinaisia analytiikan kokeiluja ei ole vielä aloitettu. (Halper & Krishnan, 2013.)

Esitason omaksunta -tasolla (pre-adoption) organisaatio kerää tietoisuutta analytiikasta osallistumalla webinaareihin ja konferensseihin. Organisaatio alkaa ymmärtämään analytiikan vaikutuksista päätöksentekoon ja liiketoimintaan ja saattaa tehdä ensimmäisiä hankintoja matalan kustannuksen teknologioihin. (Halper & Stodder, 2014.) Organisaation sisällä tiedostetaan analytiikan käyttöönotto lähitulevaisuudessa, vaikka ponnistukset analytiikan eteen ovat vielä osastokohtaisia (Halper & Krishnan, 2013).

Aikaisen omaksunnan -tasolla (early adoption) organisaatio ryhtyy käyttämään analytiikan työvälineitä sekä metodologioita. Organisaatio omaksuu datan hallintaa, raportointia ja mittaristojen käyttöä. Organisaation viettävät yleensä eniten aikaa aikaisen omaksunnan tasolla (Halper & Stodder, 2014.) Tasoon kuuluu yleensä yksi tai kaksi analytiikan konseptitodistusta (Proof of concept) jotka vakiintuvat tuotantovalmiiksi (Halper & Krishnan, 2013).

Yritystason omaksunnan (corporate adoption) saavuttaminen on merkittävä virstanpylväs organisaation analytiikan hyödyntämisessä. Tällä tasolla analytiikka saavuttaa

loppukäyttäjät ja muuttaa heidän liiketoimintatapaansa. Analyysien pohjana käytetään erilaista dataa, jopa Big Dataa. (Halper & Stodder, 2014.)

Aikaisen omaksunnan -tason ja yritystason välillä on rotko (chasm), joka kuvaa aikaa vieviä esteitä organisaation pyrkiessä laajentamaa analytiikan hyötyjä useammalle käyttäjälle eri osastoille. Esteitä aiheuttavat esimerkiksi erimielisyydet datan omistajasta sekä yhteisen vision saavuttaminen analytiikasta. (Halper & Stodder, 2014.)

Kypsyiden (mature/visionary) saavuttaneet organisaatiot toteuttavat analytiikkaan liittyviä ohjelmia sujuvasti hiotun infrastruktuurin ja datan hallintamallin avulla. Käyttäjät voivat toteuttaa datan kokeiluja ja luoda visualisointeja. Organisaatiossa valitsee aito kiinnostus data analytiikkaa kohtaan. Käytännössä kuitenkin vain muutamat organisaatiot saavuttavat kypsyysmallin ylimmän tason. (Halper & Stodder, 2014.) Gartnerin (2018) toteuttaman kyselyn mukaan vain 9 prosenttia organisaatiosta ilmoittivat saavuttaneensa ylimmän tason.

## 2.6 Datan integrointimallit

Ulkoista dataa jaetaan pääsääntöisesti kolmella tavalla; tiedostoina, katselu- ja latauspalvelun kautta ja ohjelmointirajapinnan välityksellä. Katselu- ja latauspalvelu on verkkosivu tai -palvelu, jonka kautta voi ladata tiedostot paikallisesti toiselle koneelle tai palvelimelle kriteerien valinnan jälkeen. Kriteerejä voi olla esimerkiksi tietokantataulukot tai maantieteellinen alue. (Koski ja muut, 2017.)

Katselu- ja latauspalvelussa data on yleisesti ladattavissa useassa eri tiedostoformaattissa. Aineistot on usein luokiteltu ja järjestetty kansiomaiseen tietorakenteeseen. Palvelusta valitaan ensin haluttu aineisto selaamalla tai käyttämällä hakusanaa, jonka jälkeen valitaan aineistosta haluttu osa. Lopuksi valitaan haluttu tiedostomuoto, jossa data ladataan. Katselu ja latauspalvelut ovat hyödyllisiä, kun valittavaa aineistoa on paljon, kuten

tilastoja tai paikkatietoaineistoja. Katselu ja latauspalveluita voi yleisimmin käyttää sekä käyttöliittymän että rajapinnan kautta. (Koski ja muut, 2017.)

Ohjelmointirajapinta on ohjelmiston osa, jonka avulla toteutetaan tiedonsiirtoa ohjelmistojen välillä. Ohjelmointirajapinnan kautta voidaan ladata vain tarvittu osakokonaisuus määrittelemällä tietojoukkoa rajaavia haku- ja hakuehtoja. Ohjelmointirajapinta on tehokas tapa jakaa tietoa silloin kun tietomäärä on suuri, data on reaaliaikaista tai päivittyy usein. (Koski ja muut, 2017.) Ohjelmointirajapinnasta käytetään usein englanninkielistä termiä API (Application Programming Interface).

Ohjelmointirajapinnat ovat käytettävissä joko ilman tunnistautumista, tai automaattisen rekisteröinnin avulla. Osa tietovarantojen tarjoajista edellyttää rekisteröintiä ja kerää datan hyödyntäjistä taustatietona esimerkiksi sähköpostiosoitteen. Tällöin ohjelmointirajapinnan hyödyntämiseen tarvittava avainkoodi lähetetään käyttäjän sähköpostiosoitteeseen ja rajapintaan tehtävä kysely toimii vain, jos se sisältää saadun yksilöllisen avainkoodin. (Koski ja muut, 2017.)

Datan jakelu ohjelmointirajapintojen kautta on tehokasta, mutta voi aiheuttaa kasvavia palvelin- ja tietoliikennekustannuksia kasvavan käytön myötä. Datan käyttöä voidaan hallita ohjelmointirajapinnan käytön rajoittamisella. Rajoitus voi esimerkiksi olla kyselymääräraja, mitä lataus- tai katselupalveluun voidaan lähettää vuorokaudessa. Mikäli käyttö ylittää asetetun rajoituksen, voi ylityksestä muodostua kustannuksia datan hyödyntäjälle. (Koski ja muut, 2017.) Vaihtoehtoisesti latauspalvelu voi olla maksullinen ja palvelun käytöstä peritään kohtuullinen korvaus, joka kattaa ylläpitokustannukset.

Mikäli ohjelmointirajapinnan käyttö ei vaadi rekisteröitymistä, voi palvelun käyttö olla rajoitettua kokonaiskapasiteetiltaan. Tällöin rajapinnan maksimikapasiteetti jakautuu käyttäjien kesken. Rajoitus voi esimerkiksi olla 300 kyselyä minuutissa. Rajapintaa käytettäessä on sattumanvaraista, missä vaiheessa sallittu kyselymäärä tulee täyteen. (Koski ja muut, 2017.)

### 3 Ulkoinen data

Tietojärjestelmien tuottaman datan määrä kasvaa eksponentiaalisesti. Euroopan komission arvion mukaan maailmassa tuotetun datan määrä tulee kasvamaan vuoden 2018 33 tsettatavun toteumasta 175 tsettatavuun vuonna 2025 (European Commission, 2020). Datamäärän kasvaessa kasvaa myös käytettävissä olevan datan potentiaalinen arvo (Schatsky ja muut, 2019).

Käytettävissä oleva data jaetaan organisaatioiden itse tuottamaksi sisäiseksi dataksi sekä organisaation ulkopuolelta hankittavaksi ulkoiseksi dataksi (Hartmann ja muut, 2016). Sisäisestä datasta käytetään myös termiä ensisijainen data (primary data). Ulkoisesta datasta käytetään termiä toissijainen data (secondary data). (Crossman, 2019.)

Ellram ja Tate (2016) määrittelevät ulkoiseksi dataksi määrällisen tai laadullisen datan, joka on kerätty toisen tahon toimesta ja toiseen käyttötarkoitukseen kuin mikä on datan aiottu käyttö. Crossmanin (2019) mukaan ulkoista dataa ei ole alun perin kerätty vastamaan aiotun käyttötarkoituksen kysymykseen vaan muuta käyttötarkoitusta varten. Tämä tarkoittaa sitä, että sama data voi toimia eri käyttötarkoituksissa sekä ensisijaisena datana että toissijaisena datana.

Verhoevenin (2017) mukaan on tapahtumassa muutos, jossa ulkoisen datan arvo on muuttumassa tärkeämmäksi kuin organisaatioiden sisäisten järjestelmien tuottaman datan arvo. Myös Woods (2012) jakaa saman ajatuksen. Muutoksella viitataan esimerkiksi siihen, että ulkoinen data voi tuottaa organisaatiolle sellaisia näkökulmia, joita ei aiemmin ollut saatavilla. Edellä esitetyn takia data-analyysin edelläkävijäorganisaatiot käyttävät hyväkseen saatavilla olevaa ulkoista dataa. (Schatsky ja muut, 2019.)

Forresterin (2017) tekemän tutkimuksen mukaan 92 prosenttia organisaatiosta tiedostaa kasvavan tarpeen ulkoisen datan käyttöön. Gartnerin (2018) selvityksen mukaan noin puolet organisaatiosta käyttävät ulkoista dataa osana analytiikkaa. Erään tutkimuksen

mukaan noin puolet organisaatioista ilmoittivat kaupallistavansa sisäistä dataa (Belissent, 2017).

Organisaatiot hyötyvät ulkoisen datan käytöstä monella tapaa. Crossmanin (2019) mukaan suurin hyöty on taloudellinen säästö. Ulkoisen datan hyödyntäminen muun muassa säästää organisaatioiden aikaa ja resursseja. (Johnston, 2014). Ellram ja Tate (2016) luettelevat ulkoisen datan hyödyt seuraavasti:

- Suuri saatavilla oleva datamäärä.
- Tiedon kerääminen aiheuttaa vähemmän kuluja.
- Tiedon keräämiseen menee vähemmän aikaa.
- Ulkoista dataa voidaan tarkistella objektiivisemmin kuin sisäistä dataa.
- Tiedon keräämiseen tarvitaan vähemmän henkilöresursseja.
- Datan yhdistely auttaa tarkistelemaan ilmiötä perusteellisemmin.
- Datan tulkinnaissa voidaan käyttää vakiintuneita mittauksia.
- Data on tietyissä tilanteissa auditoitua, kuten taloudelliset tiedot.

Ulkoisen data voi olla saatavilla vapaasti, vaatia datan tarjoajan lupaa tai olla maksullista (Ellram & Tatem, 2016). Tässä työssä otetaan tarkempaan tarkasteluun avoin data sekä maksullinen data, jotka käydään läpi seuraavissa luvuissa.

### **3.1 Avoin data**

Avoindata.fi (n.d.) mukaan avoimen datan juuret johtavat 1980- ja 1990-luvulla syntyneisiin vapaita ohjelmistoja ja avointa lähdekoodia (open source) edistäviin liikkeisiin. Ohjelmistoja on kehitetty avoimissa yhteisöissä alkaen 1960-luvulta, jolloin hakkerikulttuurin myötä syntyi vapaiden ohjelmistojen liike. Suljettujen ohjelmistojen merkitys alkoi kasvaa 1980-luvulla, kunnes 2000-luvulla avoimesta mallista tuli jälleen suosittu. (Vainio, 2011.)



Kilpailu avoimen ja suljetun mallin välillä on alkujaan ideologinen, avoimuuden ollessa pääsääntö erityisesti akateemisessa maailmassa. Kilpailun taustalla olevat kysymykset koskevat muun muassa informaation omistamista, käyttäjän vapautta ja yhteisöllisten arvojen turvaamista. (Vainio, 2011.) Avoimuus on tänä päivänä eräänlainen yhteiskuntaa läpileikkaava megatrendi, joka vahvistuu monilla aloilla (Avoindata.fi, n.d.).

Avoindata.fi-palvelu (n.d.) määritelmän mukaan Data on koneluettavaa tietoa, esimerkiksi taulukkoja, tekstiä, kuvia, karttoja, videoita ja äänitiedostoja. Dataan liittyvänä määritelmänä, Avoin on palvelun mukaan datan avoimen hyödyntämisen mahdollistava käyttölupa. Käyttölupa eli lisenssi on käytännössä Creative Commons BY 4.0 tai CC0 1.0. Avoinesta datasta voidaan puhua siinä tilanteessa, kun data julkaistaan uudelleenkäytön sallivalla lisenssillä maksutta ja koneluettavassa muodossa. (Avoindata.fi, n.d.)

CC BY 4.0 lisenssillä julkaistun datan käyttöä, jakelua ja muokkausta ei rajoiteta, mutta käyttäjän tulee viitata alkuperäiseen lähteeseen. CC0 1.0 lisenssillä julkaistun datan käyttöä, jakelua ja muokkausta ei rajoiteta, eikä viittausta alkuperäiseen lähteeseen tarvita. (Creative Commons, 2019.)

Helsinki Region Infoshare (HRI) -palvelun (2017a) mukaan avoin data on julkishallinnon, organisaatioiden tai yritysten tuottamaa tai niille kertynyttä julkista tietoa, joka on avattu rakenteisessa muodossa vapaasti ja maksutta kaikkien hyödynnettäväksi. Valtiovarainministeriö (n.d.) määrittelee avoimeksi dataksi julkisen organisaation tuottamaa tai hallinnoimaa tietoa, joka on konekielisessä muodossa ja kenen tahansa maksutta käytettävissä, muokattavissa ja jaettavissa sekä yksityisiin että kaupallisiin tarkoituksiin.

Julkinen data ei ole sama asia kuin avoin data. Julkiseen dataan on kaikilla pääsy ja ihmiset pääsevät lukemaan tietoja esimerkiksi verkkosivuilta mutta julkinen data ei yleensä ole avoimesti saatavissa. Kun julkinen data on avoimesti saatavissa, puhutaan avoimesta datasta, jota kansalaiset ja yritykset voivat käyttää omiin tarkoituksiinsa tasavertaisesti julkisen hallinnon kanssa. (HRI, 2017a.) Yleisestä käsityksestä poiketen avoin data ei ole

pelkästään julkista avointa dataa, vaan yrityksistä esimerkiksi lääkeyhtiöt ovat kunnostautuneet avoimen datan julkaisussa (Marjamäki, 2014).

Datan avaamista edistävä kansainvälinen organisaatio, Open Knowledge, määrittelee avoimen datan sisältävän seuraavat kriteerit (Open Definition, n.d.):

1. Julkisuus

Datan on oltava julkista tai saatavilla avoimen lisenssin nojalla. Datan mukana toimitettavat lisäehdot eivät saa olla ristiriidassa datan julkisuuden tai avointen lisenssiehtojen kanssa.

2. Koneluettavuus

Data on tarjottava avoimessa formaatissa, joka on tietokoneohjelmallisesti käytökelpoisessa ja muokattavassa muodossa.

3. Saavutettavuus

Data on oltava ladattavissa Internetistä kokonaisuutena ja maksutta tai korkeintaan kohtuullista kertakorvausta vastaan.

HRI (2017a) lisää kriteereihin vielä sen, että datan rakenne ja merkitys pitää kuvata käyttäjille ymmärrettävästi. Kuvaamiseen käytettävä metadata auttaa käyttäjää hahmottamaan datan sisällön ja tulkitsemaan dataa. Datan olemassaolon ja sijainnin tulee lisäksi olla yleisesti tunnettu. (HRI, 2017a.)

Nyrkkisääntönä kaikki julkaistavissa oleva data voidaan avata. Mutta toisaalta kaikki verkon kautta julkaistu aineisto ei välttämättä täytä avoimen datan kriteerejä. (Avoindata.fi, n.d.)

### 3.2 Avoin data Suomessa

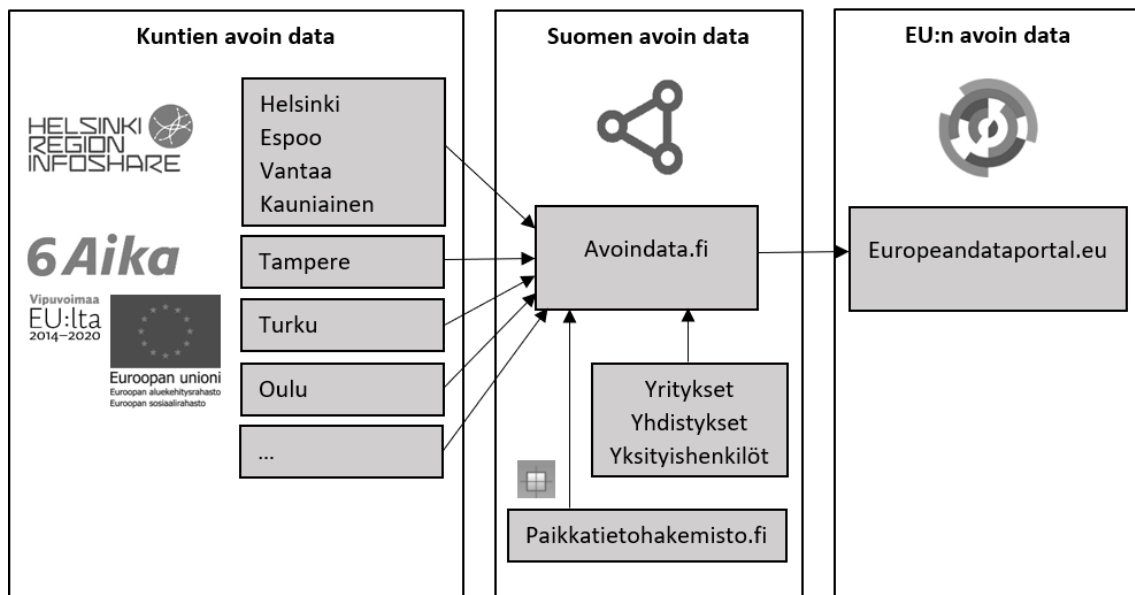
Suomessa datan avaaminen alkoi kiihtyä vuonna 2009 Euroopan unionin PSI-direktiivin sekä Britannian ja Yhdysvaltojen esimerkkien vauhdittamana (Avoindata.fi, n.d.). PSI-direktiivi (Public Sector Information) on julkisen sektorin hallussa olevien tietojen uudelleenkäyttöä koskeva EU:n direktiivi. Direktiivi uusittiin kesäkuussa 2019 ja on uudelta nimeltään Euroopan Parlamentin ja Neuvoston Direktiivi (EU) 2019/1024 avoimesta datasta ja julkisen sektorin hallussa olevien tietojen uudelleenkäytöstä. (Forsström, 2019.) Vuonna 2009 avoin data alkoi arkipäiväistymään, kun useat maat julkaisivat aloitteita avata hallinnon tietoja. Suomessa avoimen tiedon politiikan kehittäminen alkoi vuonna 2011, kun hallitus julkisti periaatepäätöksen julkishallinnon digitaalisten tietoaaineistojen saatavuudesta, jonka jälkeen hallitusohjelmaan kirjattiin julkisten tietovarantojen määrätietoinen avaaminen. (Avoindata.fi, n.d.)

Suomessa ensimmäiset merkittävät tietoaaineistot avattiin toukokuussa 2012, kun Maanmittauslaitos avasi maastotietojaan. Tätä ennen oli jo avattu lukuisia paikallistason aineistoja. (Avoindata.fi, n.d.) Vuonna 2013 Ilmatieteen laitos avasi säätiedot avoimeksi dataksi, kun tietovarantojen avaamista vauhditettiin vuosina 2013-2015 valtiovarainministeriön asettaman avoimen tiedon ohjelman toimenpitein (Kauhanen-Simanainen & Suurhasko, 2015).

EU:n avoimen datan portaali (EU ODP) perustettiin vuonna 2012 asiakirjojen uudelleenkäytöstä annetun komission päätöksen mukaisesti. Päätöksen mukaan kaikkia EU:n toimielimiä pyydetään asettamaan datansa avoimesti saataville aina kun se on mahdollista. (EU:n avoimen datan portaali, n.d.). Portaalin perustaminen pohjautuu Euroopan komission joulukuussa 2011 julkaiseeseen avoimen datan strategiaan, jonka tavoitteena on edistää tietovarantojen hyödyntämistä jäsenmaissa muun muassa tutkimusohjelmien ja infrastruktuurihankkeiden avulla. (Kauhanen-Simanainen & Suurhasko, 2015.) Avoimen datan portaali tarjoaa pääsyn EU:n toimielinten ja laitosten tuottamaan jatkuvasti kasvavaan datavarantoon. Portaalia ylläpitää EU:n julkaisutoimisto. (EU:n avoimen datan portaali, n.d.)

Kansallisen tason portaali (Avoindata.fi) avattiin Suomeen vuonna 2014. Avoindata.fi on avoimen tiedon ja yhteentoimivuuden palvelu, jonka tarkoituksena on kerätä tiedot kaikista avoimena datana tarjottavista aineistoista yhteen palveluun (Koski ja muut, 2017). Avoindata.fi-palvelu muodostaa ensisijaisesti julkisen hallinnon avoimen datan hakemiston. Palvelu on Väestörekisterikeskuksen tuottama ja säädetty valtiovarainministeriön asetuksen (607/2016) 1 §:ssä. Palvelu on tarkoitettu datan julkaisemiseen ja hyödyntämiseen, kohderyhminään muun muassa kansalaiset, yritykset ja viranomaiset. Data on julkista ja teknisesti avointa ja dataa voidaan käyttää ja hyödyntää eri tarkoituksiin sekä julkisessa hallinnossa että yksityisellä sektorilla. (Avoindata.fi, n.d.)

EU:n avoimen datan portaali kokoaa dataa jäsenmaiden avoimen datan portaaleista harvestoimalla, eli keräämällä tietoa automaattisesti. Suomesta tietoaineistot kerätään Avoindata.fi-palvelun kautta. Siirrettävät tietoaineistot ja niiden metadata käännetään koneellisesti englanninkielisiksi, minkä vuoksi käännosten laatu ja ymmärrettävyys ovat heikkoja. Osa materiaalista myös jää suomenkieliseksi. (Koski ja muut, 2017.) Tiedonkeruu on esitetty kuviossa 7.



**Kuvio 7.** Avoimen datan kerääminen Suomesta EU-tasolle (mukaillen Koski ja muut, 2017).

Avoindata.fi-palveluun kerätään julkisen hallinnon organisaatioiden tuottama avoin data (Kauhanen-Simanainen & Suurhasko, 2015). Palveluun siirretään harvestoimalla meta-data pääkaupunkiseudun Helsinki Region Infoshare (HRI) -palvelusta, Maanmittauslaitoksen ylläpitämästä paikkatietohakemistosta, Oulun dataportaalista ja Suomen ympäristökeskuksen (SYKE) avoimen tiedon palvelusta. Yritykset, yhdistykset sekä yksityishenkilöt voivat tuottaa avointa dataa ja jakaa datan Avoindata.fi-palveluun. 29.3.2020 palvelussa oli 134 dataa julkaissutta tuottajaa, joista 9 oli yrityksiä, 5 yhdistyksiä, säätiöitä tai järjestöjä sekä 38 yksityishenkilöä. (Avoindata.fi, n.d.)

EU:n avoimen datan portaalissa on yli miljoona tietoaainestoa 35 maasta (European Data Portal, 2020). EU-maiden välillä on suuria eroja kansallisten portaalien käyttöönotossa. Espanja otti ensimmäisenä maana avoimen datan portaalin käyttöön vuonna 2009. Espanja on myös yksi EU:n avoimen datan politiikan edelläkävijämaista. (Carrara ja muut, 2015.) Malta otti viimeisenä maana avoimen datan portaalin käyttöön vuoden 2019 alussa (Blank, 2019).

### **3.3 Avoimen datan jakelukanavat**

Avoimella datalla on useita eri jakelukanavia, ja sama data voi olla saatavilla useassa eri jakelukanavassa. Avoin data voi olla tarjolla datan tarjoajan omassa palvelussa tai eri tietoaainestojakoavissa portaaleissa.

Avoindata.fi on portaali, joka kokoaa yhteen Suomen avoimen datan ja koostuu pääasiassa julkisen hallinnon avoimesta tietoaainestosta. Kansallisen portaalin lisäksi, avointa dataa voidaan koota alueellisiin dataportaaaleihin. Alueellisia dataportaaaleja ovat esimerkiksi Helsinki Region Infoshare (HRI), Lounaistieto, Tampere Region Infoshare sekä Oulu open data.

HRI on Helsingin, Espoon, Vantaan ja Kauniaisten kaupunkien yhteinen avoimen datan portaali (HRI, 2017a). Lounaistieto on Lounais-Suomen alueellinen tietopalvelu, joka

ylläpitää avoimen datan portaalia (Lounaistieto.fi, n.d.). Tampere Region Infoshare on Tampereen kaupungin dataportaali, joka tarjoaa avointa dataa Tampereen seudulta (Tampere, n.d.). Oulu open data on Oulun kaupungin dataportaali, joka tarjoaa avointa dataa Oulun seudulta (Oulu, n.d.).

Alueelliset portaalit on toteutettu yhteistyössä avoimella lähdekoodilla ja palvelu on helpposti monistettavissa myös muiden alueiden tai kuntien käyttöön. Alueellisista dataportaaaleista informaatio virtaa kansalliseen Avoindata.fi palveluun. (Jokela, 2015.) Alueellisissa portaaaleissa on tarjolla alueen väestöä, taloutta, hyvinvointia, liikennettä ja matkailua koskevia tietoaaineistoja sekä runsaasti erilaisia paikkatietoaaineistoja (HRI, 2017a).

Jokela (2015) on nostanut keskustelun alueellisten avoimen datan jakelukanavien tarpeellisuudesta. Vaihtoehdoksi hän esittää, että Suomen kokoisessa maassa voisi riittää pelkkä kansallisen tason portaali avoimen datan jakeluun. Toisaalta Jokela kokee, että eritasoiset ja eri toimijoille kohdenetut portaalit eivät kilpaile keskenään, ja mahdollinen kilpailu nähtäisiin merkinä elinvoimaisesta ekosysteemistä. Kilpailun kautta nähdään lisäksi, onko lopputuloksena vain yksi elinvoimainen yksilö, vai löytävätkö eri portaalit oman ekolokeronsa.

Useat eri avoimen datan jakelukanavat saattavat aiheuttaa epäselvyyttä avoimeen dataan tutustuville organisaatioille. Epäselvyys saattaa muodostua tietoaaineiston oikeasta lähteestä ja datan ajantasaisuudesta, mikä puolestaan lannistaa organisaatioiden innokkuutta avoimen datan käyttöön. Myös Schatsky ja muut (2019) tunnistavat datan toimitajamarkkinoiden monimutkaisuuden yhdeksi haasteeksi ulkoisen datan käytölle.

Paikkatietohakemisto on valtakunnallinen palvelu, johon on koottu paikkatietoja koskevaa metadataa. Esimerkiksi Luonnonvarakeskuksen ja Suomen ympäristökeskuksen avoin data on suurelta osin paikkatietosidonnaista. Paikkatietoihin liittyy avoimen datan Paikkatietoikkuna.fi-portaali. (Koski ja muut, 2017.)

Avoimen tutkimusaineistojen julkaisualustana toimii AVAA-palvelu. Palvelu edistää eri tieteenalojen tutkimusdatan avointa saatavuutta ja on suunnattu sekä aineistojen avaa-jille että jatkohyödyntäjille. AVAA on opetus- ja kulttuuriministeriön tarjoama ja palvelua tuottaa CSC – Tieteen tietotekniikan keskus Oy. (AVAA, n.d.) Tutkimusdatan palveluita on useita muitakin, kuten Fairdata IDA, ELIXIR Suomi ja B2SHARE, mutta palvelut ovat pääosin tarjolla ainoastaan koulu- ja tutkimuslaitoksille.

Itämeri.fi on Suomen avoimen meridatan portaali, joka tarjoaa kootusti eri laitosten tuottamia meridatapalveluja. Portaalin kautta on haettavissa Itämereen liittyviä paikka-tieto- ja tutkimusaineistoja kaikkien meritietoa tuottavien laitosten tietokannoista. (Syke, 2020b.)

### **3.4 Datan avoimuuden edistäjät**

Datan avoimuutta edistävä toiminta käynnistyi Suomessa viime vuosikymmenen vaihteessa, kun joukko yksittäisiä ihmisiä käynnisti aktiivisen keskustelu aiheesta ja vuonna 2009 järjestettiin ensimmäinen avoimen datan kilpailu Apps4Democracy. Kymmenen vuoden aikana avoimen datan toimijajoukko on laajennut valtavasti niin datan avaajien kuin hyödyntäjien osalta ja kulttuurinmuutos on ollut valtava. Avoin data on vuosien varrella kokenut hypetyksensä ja on sittemmin arkipäiväistynyt osan julkishallinnollisen ta-hon normaalitoimintaa. (OKFI, 2019.)

Datan avoimuuden edistäminen on aktiivista, verkostomaista ja kansainvälistä yhteis-työtä. Suomi on ollut mukana mm. EU:n SharePSI -hankkeessa, jossa kehitetyt datan avoimuuden parhaat käytännöt on sisällytetty Suomen kansallisen tason jakeluportaaliin. Avoimen datan parhaita käytäntöjä ovat Suomessa olleet kehittämässä julkishallinnon dataa avaavat virastot, Suomen Kuntaliitto, Open Knowledge Finland ry ja 6Aika-hanke. (Valtiovarainministeriö, n.d.)

Open Knowledge Finland (OKFI) on rekisteröity yleishyödyllinen yhdistys, joka edistää tiedon avaamista, avoimen tiedon hyödyntämistä sekä avoimen yhteiskunnan kehittämistä Suomessa. OKFI on perustettu vuonna 2012 ja yhdistyksellä on yli 500 jäsentä, jotka koostuvat organisaatioista, yrityksistä sekä yksityisistä henkilöistä. Yhdistys on aktiivinen toimija ja järjestää avoimen datan koulutuksia, kilpailuja sekä on mukana monissa muiden koordinoimissa avoimen datan projekteissa. OKFI oli mukana laatimassa Suomen julkisen hallinnon suositusta JHS 189, Avoimen tietoaineiston käyttöluupa. (OKFI, n.d.)

6Aika-strategia on vuonna 2014 käynnistetty hanke, jossa kuuluvat Helsinki, Espoo, Vantaa, Tampere, Turku ja Oulu. Strategian hankkeissa kaupungit ovat luoneet datan avaamiselle yhteisiä toimintamalleja, avanneet harmonisoituja ohjelmointirajapintoja ja tukeneet avoimen datan hyödyntämistä liiketoiminnassa. (6Aika, n.d.)

Edellä mainittujen toimijoiden lisäksi Suomessa on merkittävä määrä muita organisaatioita ja palveluita, jotka edistävät datan avoimuutta. Avoin Satakunta edistää datan avoimuutta Satakunnan alueella ja Facebookissa toimii Finnish Open Data Ecosystem -ryhmä, joka sisältää yli 4000 jäsentä. Jäsenistö koostuu julkishallinnon ja yritysten edustajista sekä yksityisistä kansalaisaktivisteista.

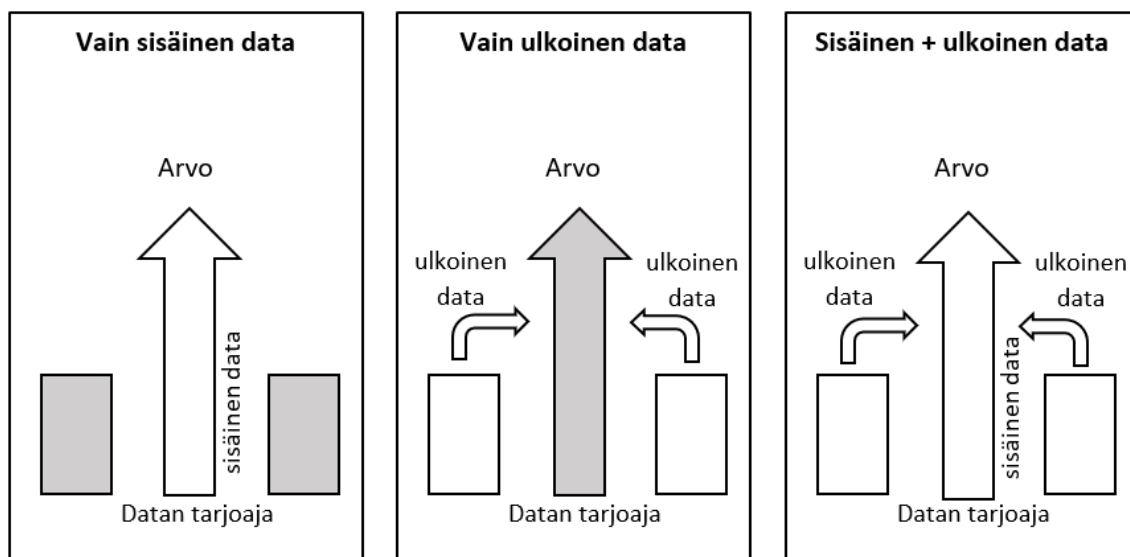
Datan avoimuuden edistämiseksi järjestetään myös avoimen datan kilpailuita, joissa haetaan uusia avoimen datan käyttötapoja. Kilpailuita on järjestetty vuodesta 2009 alkaen Apps4Finland-nimellä. Vuodesta 2015 alkaen kilpailu nimi muuttui Open Finland Challengeksi. Kilpailua ovat järjestäneet Forum Virium Helsinki ja Open Knowledge Finland.

### **3.5 Maksullinen data**

Maksullinen data on normaalisti organisaation tarjoamaa sisäistä dataa, joka ei ole muille avoimesti saatavilla. Maksullinen data voi olla myös useista eri ulkopuolista datalähteistä koottua dataa, jota organisaatio tarjoaa edelleen muille. Organisaatio tuottaa



dataan lisäarvoa muuntamalla datan informaatioksi analytiikan avulla. Kolmas vaihtoehto on se, että organisaatio yhdistää omaan sisäiseen dataansa ulkopuolista dataa, ja tarjoaa tällä tavalla rikastettua dataansa muille (ks. kuvio 8). (Verhoeven, 2017.)



**Kuvio 8.** Maksullisen datan vaihtoehdot (mukaillen Verhoeven, 2017).

Davisin (2019) mukaan maksullisen datan tarjoaminen on vielä suhteellisen uusi käytäntö. Organisaatiot ostavat dataa mutta kaupankäyntiin ei ole olemassa yhä suurta luotettavaa markkinapaikkaa, jossa data olisi laadukkaaksi seulottua. Vastaavasti datan myyjille ei ole tarjolla pankkitoimintaa vastaavaa selvityspalvelua (clearing), joka takaisi datan laadun ja auttaisi ostajan löytämisessä. (Davis, 2019.)

Datan hankinnassa on vaihtoehtoina joko ostaa pelkkää dataa tai lisäksi datasta muodostettuja näkemyksiä. Schatsky ja muut (2019) luokittevat saatavilla olevat datapalvelut yksinkertaiseksi palveluksi (simple data services), älykkääksi palveluksi (smart data services) ja adaptiiviseksi palveluksi (adaptive data services). Yksinkertaisessa palvelussa datan tarjoajat keräävät dataa useasta eri ulkoisesta lähteestä ja myyvät datan puhdistetussa muodossa. Älykkäässä palvelussa datan voi ostaa valmiiksi analysoidussa muodossa, esimerkiksi pisteytettynä tai muulla tavoin merkittynä. Adaptiivisessa palvelussa

organisaatiot voivat toimittaa omaa dataansa analysoitavaksi. Palveluntarjoajat yhdistävät toimitetun data muihin datalähteisiin lisäinformaation tuottamiseksi. (Schatsky ja muut, 2019.)

Maksullisten tietojen maailmanmarkkinoilla on tällä hetkellä paljon liikehdintää ja markkinoille on tulossa uusia asiantuntijapalveluiden tarjoajia (Davis, 2019). Datan tarjoajat ovat voineet erikoistua tietyille aloille kuten terveystalouteen tai valmistavaan teollisuuteen. Palveluntarjoajat myös yhdistelevät julkista avointa dataa ja kumppanidataa muodostaakseen uusia datavirtoja. (Schatsky ja muut, 2019.)

Organisaatioiden on oltava tietoisia säännöistä, jotka koskevat ostettavaa tai myytävää dataa. Mikäli data on esimerkiksi asiakkaan luomaa, on oltava tietoinen kenelle datan oikeudet kuuluvat. (Davis, 2019.)

## 4 Ulkoisen datan tarjoajat Suomessa

Valtiovetoinen datan avaaminen näkyy vielä tämänhetkisessä avoimen datan tuottajajoukossa. Suurin osa avoimen datan tuottajista koostuu julkishallinnon organisaatioista, valtion viranomaisista tai julkisyhteisöjen omistamista yrityksistä. Useat kunnat ja kaupungit ovat aktivoituneet datan avaamisessa ja tarjoavat avoimena datana myös muuta tietoaaineistoa kuin julkishallinnon dataa. Yksityisten yritysten avoimen datan tarjoama on vielä pieni. Osa yrityksistä tarjoaa kuitenkin myös avointa dataa maksullisen datan lisäksi.

### 4.1 Avoimen datan tuottajat

Lista merkittävimmistä avoimen datan tuottajista sisältää seuraavat toimijat (mukaillen DataBusiness.fi, n.d.):

- Maanmittauslaitos
- Geologian tutkimuskeskus
- Ilmatieteenlaitos
- Tilastokeskus
- Väylä (ent. Liikennevirasto)
- Traffic Management Finland Oy
- Traficom (Liikenne- ja viestintävirasto)
- ITS Factory
- Terveyden ja hyvinvoinnin laitos
- SYKE (Suomen ympäristökeskus)
- Ruokavirasto (ent. Elintarviketurvallisuusvirasto, Evira)
- Tukes (Turvallisuus- ja kemikaalivirasto)
- Luke (Luonnonvarakeskus)
- Suomen Pankki ja Finanssivalvonta
- Patentti- ja rekisterihallitus
- Fingrid Oyj

- Helen Oy
- Teknologian Tutkimuskeskus VTT Oy
- HSL
- HSY
- Nysse
- Visit Finland

Maanmittauslaitos tarjoaa avoimien aineistojen tiedostopalvelun kautta avoimena datana kartta- ja ilmakuva-aineistoja. Karttapaikka-palvelun kautta pääsee selaamaan maasto- ja taustakarttoja sekä kiinteistöjaotuksia. Kiinteistötietopalvelussa voi selata valtakunnallisen kiinteistötietojärjestelmän (KTJ) tietoja. Osoitehaavi sisältää tietoa sisäänkäyntien ja kulkupisteiden sijainnista ja ominaisuuksista. Osoitepalvelun tiedot kerätään joukkoistamalla. (Maanmittauslaitos, n.d.)

Maanmittauslaitoksen tarjoama Paikkatietoikkuna on kansallinen paikkatietoportaali, joka esittelee paikkatietoaineistoja ja -palveluja sekä niiden hyödyntämismahdollisuuksia. Suomigrammi on maastotietojen tilastotyökalu, joka esittää muun muassa lomasuntojen sijoittumista ja havainnollistaa Suomen järvisimmät kunnat. Maanmittauslaitos tarjoaa lisäksi tilastotietoa kiinteistökaupoista Tiedot koskevat asuinpientalo-, maa- ja metsätalous- sekä lomakiinteistöjä. (Maanmittauslaitos, n.d.)

Geologian tutkimuskeskus tarjoaa rajapintapalvelunsa kautta avoimena datana geofyysiikkaan-, kallioperään-, kiviainesvarantoihin-, maaperään- sekä pohjatutkimuksiin liittyvää tietoa-aineistoa. Hakku on haku- ja latauspalvelu geologisille tietotuotteille, joka sisältää julkaisuja, raportteja, karttoja, paikkatietotuotteita, valokuvia ja karttapiirroksia. Karttapalvelun kautta on haettavissa havainto- ja mittaustietoa metallimalmeista, teollisuusmineraaleista ja malminetsinnästä, kallioperätutkimuksesta, vesihuollosta, maa- ja kallioperästä, rakentamisesta ja maankäytön suunnittelusta sekä turvemaiden kartoituksesta. (GTK, n.d.).

Ilmatieteenlaitoksen sivuilla avoimen datan tietoaaineistot voidaan jakaa reaaliaikaisiin tietoihin, aikasarjoihin sekä ennustetietoihin. Aineistot kattavat sää-, meri- ja ilmastohavaintoja, tutka- ja salamahavaintoja sekä kansallisen sääennustemallin tiedot. (Ilmatieteen laitos, 2019.)

Tilastokeskus tarjoaa avoimena datana tilastoaineistoa eri aihealueittain koko Suomesta, kunnittain, postinumeroalueittain tai karttaruuduittain. Tilastokeskuksen avoimet tietoaaineistot sisältävät StatFin-tilastotietokannan, kuntien avainluvut, kuntien ja kuntayhtymien raportoimat taloustiedot, maahanmuuttajat ja kotouttamisen -tietokannan, koto-tietokannan, väestönlaskentatietoja, Paavo-palvelun kautta tarjottavan postinumeroaluettaisen tilastotiedon, toimipaikkalaskurin, ympäristötilinpidon avainluvut, tieliikenneonnettomuudet, verohallinnon -ja traficomin tilastotietokannan, tilastopalvelu Rudolfin, joka tarjoaa matkailun aihepiiriin kuuluvia tietoja sekä muutamia kansainvälisiä tietokantoja. (Tilastokeskus, n.d.)

Väylä tarjoaa sivullaan avoimen datan aineistona tieverkkoja koskevaa aineistoa. Tieverkko kattaa tierekisterin, digiroadin, tieliikenneonnettomuuksien datan ja liikenne-merkkien kirjasintyyppidatan. Tierekisteri sisältää tietoja Väylän vastuulla olevista maanteistä ja teiden liikenteestä. Digiroad on kansallinen tietietojärjestelmä, johon on koottu koko Suomen tie- ja katuverkon keskilinjageometria sekä tärkeimmät ominaisuustiedot. (Väylä, 2019.)

Traficomin avoimen datan verkkopalvelussa on saatavilla liikenteeseen ja viestintään liittyviä tietoaaineistoja sekä paikkatietoaaineistoja. Liikenteen tietoaaineistot sisältävät Traficomin ylläpitämiin rekistereihin rekisteröidyt tiedot rautateiden kalustosta, ilma-aluksista, ajoneuvoista, aluksista ja vesikulkuneuvoista. Viestinnän tietoaaineistot sisältävät tietoja matkaviestinverkosta, kiinteästä verkosta sekä radio- ja TV-aseamista. Paikkatietoaaineistot sisältävät muun muassa merikartoituksen syvyysmittaustietoja. (Traficom, 2020.)

Traffic Management Finland Oy tarjoaa Digitraffic-palvelussa avointa dataa tie-, rata- ja vesiliikenteestä. Tieliikenteen tietoaineistot sisältävät ajantasaista liikenne- ja olosuhdetietoa. Rataliikenteen tietoaineistot sisältävät junien aikataulut, toteumatiedot, sijainnit ja kokoonpanot. Vesiliikenteen tietoaineistot sisältävät merivaroitukset, alusten sijaintitiedot ja Portnet -tiedot alusten satamakäynneistä. (Traffic Management Finland, n.d.)

ITS Factory on innovaatio-, kokeilu- ja kehitysalusta, jossa yritykset ja yksittäiset kehittäjät voivat kehittää, testata ja tuottaa älykkäitä liikennetkaisuja. ITS (Intelligent Traffic Systems and Services) Factorya johtaa Tampereen kaupunki yhteistyössä alan yritysten kanssa. ITS Factoryn avoin data kattaa Tampereen alueen joukkoliikenteeseen liittyviä tietoaineistoja. (ITS Factory, 2020; Business Tampere, n.d.)

Terveyden ja hyvinvoinnin laitoksen avoin data kattaa sosiaali- ja terveysalan tietoaineistoja. Tietoaineistot sisältävät muun muassa sairaanhoidon toimenpiteet ja hoitoon pääsyn jonotilanteet, terveydenhuollon käynnit ja käyntisyyt alueittain ja eri ajanjaksoittain, infektioautien alueellinen ja ajallinen esiintyminen, tilastotietoja suomalaisten terveydestä ja hyvinvoinnista (Sotkanet), vertailutietoa kuntien terveydenedistämisaktiivisuudesta (TEAvisari) sekä elintarvikkeiden koostumistiedot (Fineli). (THL, 2020.)

SYKE tarjoaa avointa dataa kestävästä ympäristöstä ja yhteiskunnan rakentamiseen. Tietoaineistoa on saatavilla pinta- ja pohjavesistä, Itämerestä, ympäristön kuormituksesta ja häiriötekijöistä, arvokkaista luonnonympäristöistä, maanpeitteestä ja rakennetusta ympäristöstä. (SYKE, 2020a.)

Ruokavirasto tarjoaa sivuillaan avointa dataa ylläpitämiensä rekisterien tietoaineistoista. Rekisterit kattavat maataloutta ja maaseudun kehittämiseen liittyvää tietoa. Ruokavirasto tarjoaa lisäksi analytiikkaportaalisissaan avointa dataa elintarvikkeiden turvallisuudesta ja laadusta sekä kasvien ja eläinten terveydestä. (Ruokavirasto, 2019.)

Tukesin avoin data kattaa tietoaineisot vaarallisista tuotteista, joiden myyntiä ja markkinointia Tukes on rajoittanut hallintopäätöksillä sekä tiedot tuotteista, joiden myyntiä ja markkinointia yritykset itse vapaaehtoisesti ovat rajoittaneet. Tietoaineistot koostuvat Tukesin ylläpitämästä markkina- ja turvallisuusrekisteristä. (Tukes, n.d.)

Luke tuottaa avointa data osana tutkimus- ja seurantatyötään. Lukesin tilastotietokannan kautta on saatavilla tietoaineistoa koskien maataloustilastoja, metsätilastoja, kala- ja riistatilastoja, erilaisia indikaattoreita sekä biotalouslaskelmia. (Luke, n.d.)

Suomen Pankki ja Finanssivalvonta ovat avanneet yhteisen avoimen datan palvelun, josta tulee löytymään Suomen Pankin ja Finanssivalvonnan osastojen julkaisemia tietoaineistoja. Ensimmäisessä vaiheessa palvelussa on ainoastaan Suomen Pankin tuottamia tilastoaineistoja. (Suomen Pankki, n.d.) 8.5.2020 Suomen Pankin ja Finanssivalvonnan avoimen datan portaali sisälsi kolme tietoaineistoa. Tietoaineistoja olivat Suomen kansalliset arkipyhät, jolloin pankit ovat kiinni, suomalaisten lainat ja talletukset sekä Euroopan pankin viitekorot.

Patentti- ja rekisterihallituksen (PRH) avoimen datan verkkopalvelun kautta on saatavilla kaupparekisterin kuulutustiedot sekä YTJ-tiedot. Kuulutustiedot sisältävät tietoa rekisteröidyistä uusista yrityksistä. Kuulutustietona on myös yritysten tietojen muuttuminen. PRH:n ja Verohallinnon yhteisen YTJ-palvelun tietoja ovat kaupparekisteriin merkityt yritysten perustiedot sekä tietoa muista rekistereistä joihin yritys on merkitty, kuten ennakkoperintärekisteri, työnantajarekisteri, arvonlisäverovelvollisten rekisteri, vakuutusmaksuverovelvollisten rekisteri ja verohallinnon asiakasrekisteri. (PRH, n.d.)

Fingrid Oyj on Suomen kantaverkkoyhtiö, joka avasi vuonna 2017 ensimmäisenä kantaverkkoyhtiönä Euroopassa avoimen datan palvelun. Yhtiön tarjoaman avoimen datan tietoaineistot sisältävät pääasiassa sähkömarkkinoihin ja voimajärjestelmään liittyvää mittaus- ja ennustetietoa. Fingrid on julkinen osakeyhtiö, jonka omistuksesta suurin osa on julkisyhteisöjen hallussa. (Leinonen, 2019.)

Helen Oy tarjoaa avoimena datana aurinkosähkön tuotannon tehotietoja reaaliaikaisesti kahdesta aurinkovoimalastaan Helsingin alueella. Historiatietona on tarjolla Helsingin kaukolämmön tehotiedot vuodelta 2016. (Helen Oy, 2020.)

Teknologian Tutkimuskeskus VTT Oy:n ylläpitämässä LIPASTO-järjestelmässä on tarjolla avointa dataa Suomen liikenteen pakokaasupäästöistä ja energiankulutuksesta. Tietoaineistoa on saatavilla tie-, raide-, vesi- ja ilmaliikenteestä sekä työkoneista. Liikennevälineiden yksikköpäästötietokanta sisältää liikennevälineiden ja työkoneiden päästökertoimet. Liikenteen päästöinventaarior sisältää liikennemuotoikohtaisilla alueille lasketut Suomen liikenteen vuotuiset kokonaispäästöt ja energiankulutus. (LIPASTO, n.d.)

HSL (Helsingin seudun liikenne) on kuntayhtymä, jonka jäseniä ovat Helsinki, Espoo, Vantaa, Kauniainen, Kerava, Kirkkonummi ja Sipoo. HSL tarjoaa Helsingin seudun joukkoliikenteeseen liittyvää avointa dataa. Reittioppaan rajapinnat tarjoavat reititykseen, geokoodaukseen, kartta-aineistoon sekä paikannukseen liittyvää dataa. Joukkoliikennerekisterin data sisältää voimassa olevat joukkoliikennereitit sekä reittien aikataulut. GTFS-RT-rajapinnat tarjoavat reaaliaikaista dataa poikkeustilanteista, pysäkkiennusteista ja ajoneuvojen sijainneista. Kaupunkipyöräasemien Origin-Destination (OD)-data sisältää tietoaineistoa Helsingin ja Espoon kaupunkipyörillä poljetuista matkoista. Beacon -rajapinta sisältää tietoaineistoa julkisen liikenteen seuraamiseksi asennetuista lähettimistä. HSL News rajapinta tarjoaa tietoaineistona hsl.fi -sivustolla julkaistut uutiset ja liikennetiedotteet. Liityntäpysäköinnin rajapinta tarjoaa tietoaineistoa Helsingin seudun liityntäpysäköinnistä. Louhin-palvelu sisältää dataa asiakastytyvyydestä ja lauttaliikenteen matkustajamääristä. Paikkatietoaineistojen latauspalvelu tarjoaa useita erilaisia HSL:n joukkoliikenteeseen tietoaineistoja, kuten pysäkit, linjat, terminaalit, nousijamäärät, kaupunkipyöräasemat ja pyörälaskennat. (HSL, n.d.)

HSY (Helsingin seudun ympäristöpalvelu) on Espoon, Helsingin, Kauniaisten ja Vantaan kuntayhtymä, jonka tehtävänä on tuottaa vesi- ja jätehuoltoja sekä ajantasaista seudullista tietoa. HSY tarjoaa avoimena datana tietoaineistoja ilmanlaadusta,



energiankulutuksesta, kasvihuonekaasupäästöistä, vesi- ja jätehuollosta sekä asumisesta ja maankäytöstä. Asuminen ja maankäyttö sisältää muun muassa karttaruutukohtaista tietoaineistoa väestön lukumäärästä, ikäjakaumasta, asumisväljyydestä, rakennusten lukumäärästä, kerrosalasta, käyttötarkoituksesta ja aluetehokkuudesta sekä aurinkopaneeleille sopivat sijainnit ja rakennuksittain lasketun aurinkosähkön tuottopotentialin. (HSY, 2019.)

Nysse tarjoaa avointa dataa Tampereen seudun joukkoliikenteestä. Reittioppaan rajapinnat tarjoavat reitteihin, pysäkkeihin, aikatauluihin ja reititykseen liittyvää tietoaineistoa. GTFS-RT-rajapinnat tarjoavat reaaliaikaista tietoaineistoa ajoneuvojen sijainneista, pysäkkikohtaisista autojen saapumisennusteista sekä häiriötiedotteista. (Nysse, 2020.)

Visit Finland tuottaa tilasto- ja tutkimustietoa Suomeen suuntautuvasta matkailusta yritysten ja matkailuelinkeinon toiminnan suunnittelun ja päätöksenteon tueksi. Visit Finland tarjoaa Suomen matkailualan avointa dataa tilastopalvelu Rudolfista. (Business Finland, n.d.)

## **4.2 Maksullisen datan tuottajat**

Suurta osaa maailman datasta hallitsee pieni joukko suuria teknologiayrityksiä (European Commission, 2020). Suurimpia datan hallitsijayrityksiä ovat Facebook, Google, Tencent, Baidu ja Alibaba (Sun & Huo, 2019).

Suurimmista yrityksistä Facebook ja Google tarjoavat sisäistä dataansa rajatusti muille. Facebookin Analytics-analytiikkatyökalulla voi selvittää miten käyttäjät ovat vuorovaikutuksessa verkkosivun, Facebook-sivun tai muiden tuettujen tapahtumalähteiden kanssa. Vuorovaikutustietojen lisäksi käyttäjistä tarjotaan demografiatiedot, kuten ikä ja sukupuoli. (Facebook, n.d.)

Google Trends tarjoaa dataa Googlelle lähetetyistä hakupyynnöistä ja kuvaa ihmisten päivittäisiä hakuja. Data on pääosin suodattamaton otos oikeista hakupyynnöistä. Data anonymisoidaan, luokitellaan ja kootaan yhteen, minkä ansiosta tulokset voidaan esittää maailmanlaajuisesti tai vastaavasti kaupunkikohtaisesti. Tarjolla on reaaliaikaista dataa sekä historiadataa vuodesta 2004 alkaen. (Google, 2020.)

Googlen ja Facebookin tavoin myös muut sosiaalisen median lähteet ovat yleisesti ulkomaalaisia yrityksiä mutta sisältävät myös Suomea koskevaa dataa. Sosiaalisen median sisältöjen kerääminen järjestelmäntarjoajilta on useimmiten maksutonta ohjelmointirajapintojen (Application Programming Interface, API) kautta. Ohjelmointirajapinta mahdollistaa kontrolloidun pääsyn sosiaalisten medioiden alustojen toiminnoille ja datalle. Vaihtoehtoinen tapa kerätä dataa on tehdä sopimus valitun sosiaalisen median alustan kanssa, joka tarjoavat datan arkistointipalvelua. Toinen sosiaalisen median lähde on datan jälleenmyyjät. Jälleenmyyjä ovat tyypillisesti yhtiöt, jotka tarjoavat APIen pohjalta kerättyyn dataan pohjaavia palveluita ja tuotteita. (Fält, 2016.)

Gnip toimii tiettyjen sosiaalisten medioiden alustojen virallisena jälleenmyyjänä ja tarjoaa pääsyn sellaiseen dataan, jota ei ole mahdollista kerätä suoraan sosiaalisen median alustojen kautta. Gnipin kautta on esimerkiksi mahdollista saada 100 prosentin pääsy Twitterin viesteihin. Gnipiä vastaava sosiaalisen median datalähde on DataSift, joka tarjoaa eksklusiivisen pääsyn Facebookin dataan. (Day Thomson, 2016.)

Suomen toiminta-aluetta koskevan maksullisen datan tuottajia ovat esimerkiksi seuraavat yritykset ja julkisen hallinnon organisaatiot:

- Bisnode Finland Oy
- Suomen Asiakastieto Oy
- Profinder Oy
- Intrum Justitia Oy
- Alma Talent Oy
- Lowell Suomi Oy

- Fonecta Oy
- Posti Oy
- Tilastokeskus
- Telia Company AB
- Geomatic A/S

Bisnode Finland tarjoaa datapalveluna kuluttaja-, asiakas- ja yritystietojen päivitystä ja asiakasdatan laadun analyysiä, ajoneuvotietoja sekä kiinteistötietoja. Yritys tarjoaa lisäksi analytiikkaan pohjautuvaa palvelua myynnin, markkinoinnin ja yritystoiminnan strategiseen päätöksentekoon. (Bisnode, n.d.)

Bisnode tarjoaa Kiinteistöt-palvelussa tietoaineistot yli 1,5 miljoonasta rakennuksesta. Palvelusta on saatavilla Maanmittauslaitoksen kiinteistöihin liittyvät asiakirjat, kuten vuokraoikeustodistus, kiinteistörekisteriote, kiinteistörekisterin karttaote ja muodostamisketju taaksepäin ja muodostamisketju eteenpäin. Palvelu sisältää myös tilastotietoon perustuvan alueanalyysin, joka muodostaa objektiivisen näkemyksen alueesta, jolla kiinteistö sijaitsee. (Bisnode, n.d.)

Bisnoden ajoneuvodata kattaa Traficommin ajoneuvorekisterin tietoaineuksen sekä ajoneuvomarkkinan kehittymisen ajoneuvoluokittain. Ajoneuvodata sisältää myös jokaisen henkilöauton reaaliaikaisen arvon euroissa. (Bisnode, n.d.)

Suomen Asiakastieto tarjoaa datapalveluna asiakkaiden luottotietoja ja luottoluokituksia, asiakasrekistereiden päivitystä sekä kiinteistö- ja huoneistotietoja. Yritys tarjoaa myös tietopalveluja riskienhallintaan, taloushallintoon sekä myyntiin ja markkinointiin. (Asiakastieto, n.d.)

Suomen Asiakastiedon kiinteistö- ja huoneistotiedot sisältävät tietoaineistot kiinteistöistä, huoneistoista ja asukkaista. Kiinteistötietopalvelut sisältävät muun muassa kiinteistötietodokumentit, karttapalvelut, kauppahintarekisterin tiedot, alueen koulut,

kaupat, julkiset palvelut ja liikenneyhteydet. Asiakastiedot sisältävät tietoaineistoja rakennuksista, rakennuksessa asuvista ihmisistä tai yrityksistä, asukaskyselyistä sekä asunto-osakeyhtiöistä. (Asiakastieto, n.d.)

Profiler Oy tarjoaa datapalveluna kuluttajatietojen ja yritystietojen päivytystä ja rikastamista. Yritys tarjoaa lisäksi asiakkaiden tilasto- ja tutkimusanalyysipalvelua myyntiin, markkinointiin ja viestintään. (Profiler, n.d.)

Intrum Justitia Oy tarjoaa informaatio- ja analytiikkapalvelua yritysten ja yksityishenkilöiden luottotietojen tarkistamiseen sekä yritysten maksukyvyyn valvontaan. Palvelussa yrityksen sisäistä dataa rikastetaan Intrumin perintätiedoilla. (Intrum, 2019.)

Alma Talent Oy:n Tietopalvelut tarjoavat data- ja analytiikkapalveluita yrityksistä, henkilöistä ja kiinteistöistä. Yritystietopalvelu yhdistää yritysten rekisteritiedon, tilinpäätöstiedon, Alma Talentin journalistisen sisällön sekä sosiaalisen median viestit tarkasteltavaksi kokonaisuudeksi. Kiinteistötietopalvelut sisältävät viralliset kiinteistötiedot, kiinteistöjen kauppahinnat, kiinteistöjen karttapalvelun ja taloyhtiöiden asukkaiden tiedot. Henkilötietopalvelut sisältävät yksityishenkilöiden tiedot, kuten osoitteet, maksuhäiriöt ja liiketoimintakiellot. (Alma Talent, 2020.)

Lowell Suomi Oy tarjoaa reaaliaikaista dataa kuluttajien ja yritysten maksukyvyyn ennustamiseen. Luottotietopalvelusta on saatavilla kuluttajien ja yritysten perustiedot, mahdolliset maksuhäiriötiedot sekä tietoa maksukyvyyn muutoksista. (Lowell, 2020.)

Fonecta Oy tarjoaa analytiikkapalvelussaan dataratkaisuja yritysten verkkosivukävijöiden analysointiin. Maksuton Audience Insight -työkalu tuottaa tietoaineistoa verkkosivukävijöistä. Tietoaineistoa tuotetaan esimerkiksi verkkosivukävijöiden arvoista ja suhtautumisesta eri asioihin, kuten matkailuun ja terveyteen. Maksullinen Snoobi Analytiikka kertoo tietoa verkkosivuliikenteestä. Analytiikka kertoo mistä sivustolle tullaan, mitä sivustolta haetaan ja mistä organisaatiosta sivustolla vieraillee kävijöitä. (Fonecta, n.d.)

Posti Oy:n tarjoama maksullinen data jaetaan asiakasrekisterin ylläpidon dataan sekä muihin datapalveluihin. Asiakasrekisterin ylläpidon dataa on mahdollista käyttää asiakasrekisterin päivitykseen, tarkistamiseen ja rikastamiseen. Muita datapalveluita ovat asiakastietojen analysointi, kohderyhmäperusteisesti koottu data sekä katudata. Asiakastietojen analysointi sisältää profiloinnin, pisteytysmallit, mallinnetut rikasteet ja jakelualueiden optimoinnin. Kohderyhmäperusteisesti koottava data voi sisältää tietoaaineistoa esimerkiksi lapsiperheistä, senioreista, rakentajista, muuttajista tai ajoneuvojen omistajista. Postin omia maksullisen datan lähteitä ovat Postin osoitetietojärjestelmä (OTS), Postin sähköisten palveluiden asiakkaat -rekisteri sekä Postin keräämä evästedata, muuttajadata ja kieltoluukkudata (Posti, n.d.)

Posti tarjoaa tie- ja ympäristödataa Katudata-tilannekuvapalveluna. Palvelulla tarjotaan ratkaisua teiden ja ympäristöinfraan tehokkaampaan kunnossapitoon. Palvelun avulla voidaan tarjota dataa muun muassa liikennemerkeistä, ajoratamaalauksista, teiden pintavaurioista, tieolosuhteista ja tiesäästä. Palvelua voidaan laajentaa muihin jakelureittien varrella oleviin kohteisiin ja dataa voidaan kerätä esimerkiksi katuvalojen tai sähkökaapipien ulkoisesta kunnosta. Datan keruu ja analysointi perustuu Vaisalan RoadAI-tilannekuvapalveluun ja tiedonkeruu tapahtuu jakeluautoihin kytkettyjen kameroiden ja anturien avulla. Dataa tuottavat videokuvaa analysoiva konenäkö, Vaisalan sääsensorit sekä Postin henkilökunta. Henkilökunta raportoi havaintonsa järjestelmään. (Posti, 2018.)

Tilastokeskus tarjoaa maksullisena datana tietokantoja, joissa on tarjolla pidempiä aikasarjoja sekä yksityiskohtaisempaa tietoa kuin Tilastokeskuksen avoimen datan palveluissa. Maksullisia tietokantoja ovat aikasarjatietokanta Astika, joka sisältää suhdanne- ja taloustilastoja, tilastokatsaus, kaupunki- ja seutuindikaattorit, maaseutuindikaattorit, rakennettu ympäristö, toimialoittainen yritystietopalvelu, sijoittumispalvelu sekä väestötietopalvelu. (Tilastokeskus, n.d.)

Tilastokeskuksen tarjoama Ruututietokanta on maksullinen tuote, joka sisältää tilastoaineistoja jaettuna Suomen karttaruutuihin. Karttaruudut tarjoavat mahdollisuuden

tarkastella alueellisia ilmiöitä hallinnollisista aluerajoista riippumatta. Data sisältää tietoaineistoa alueen asukasrakenteesta, koulutuksesta, asumisesta, tuloista ja työpaikoista. Data on saatavissa koko Suomen alueelta sekä 250 m x 250 m-, 1 km x 1 km- tai 5 km x 5 km -karttaruuduittain. Tuote mahdollista myös alue- tai tietoryhmäpoiminnat. (Tilastokeskus, n.d.)

Telia Company AB tarjoaa Crowd Insights -palvelussaan dataa ihmisjoukkojen liikkumisesta. Data muodostetaan mobiililaitteiden anonyymeista sijaintitiedoista. Palvelu jakautuu eri analyysivaihtoehtoihin. Matka-analyysi kuvaa mistä ihmisjoukot tulevat ja minne ne menevät. Matka-analyysin tasoja ovat kuntataso, postinumerotaso ja Mobile Traffic Cell (MTC) -ruudukkotaso. MTC-ruudun koko on pienimmillään 500 m x 500 m kaupunkien keskustoissa. Haja-asutusalueella MTC-ruutu on suurimmillaan 4000 m x 4000 m. Sijaintianalyysi kuvaa kävijöiden määrän tietyllä alueella ja tiettyinä ajanhetkenä. Ihmisvirta-analyysi kuvaa missä ja mihin aikaan väkijoukot liikkuvat. Ihmisvirta-analyysissä MTC-ruutu on pienimmillään 250 m x 250 m. Analyysi tuottaa tietoaineistoa ihmisvirran määrästä solun läpi tiettyinä ajanhetkenä. Haluttaessa palvelu sisältää pelkästään datan, ilman analyysiä. (Telial, 2018.)

Geomatic A/S tarjoaa conzoom®-tuotteessaan maksullista dataa talouksien sosioekonomisesta asemasta, kuluttajakäyttäytymisestä ja asenteista. Suomalaiset asuinalueet jaetaan segmentointityökalulla kymmeneen eri ryhmään ja 31 luokkaan. Luokituksessa käytetään Tilastokeskuksen karttaruutupohjaisia tilastoja sekä TNS Gallupin arvo- ja kulutusikäyttäytymistietoja, jolla luokitukseseen sisällytetään kuluttajien mielipiteitä ja asenteita. conzoom®-dataa ja muita demografisia tietoja voidaan tarkastella Geomatic Audience -verkkopalvelussa 100 m x 100 m -karttaruuduittain. (conzoom®, n.d.)

## 5 Tutkimusmenetelmä

Työn tutkimusmenetelmänä käytetään kartoittavaa ekstensiivistä tapaustutkimusta. Tapaustutkimuksen nimikkeellä tehdään tutkimusta useilla tieteenaloilla, monenlaisista lähtökohdista ja erilaisin tavoittein. Tapaustutkimuksessa tutkijan ei tarvitse sitoutua vain yhteen tieteenfilosofiseen paradigmaan. Tutkimus on usein pragmaattista ja siinä hyödynnetään sekä laadullista että määrällistä aineistoa ja erilaisia analyysitapoja, minkä vuoksi tapaustutkimuksella voi olla läheinen suhde monimenetelmälliseen tutkimukseen (Eriksson & Koistinen, 2014.)

Erikssonin ja Koistisen (2014) mukaan lähestymistavaksi kannattaa valita tapaustutkimus, kun jokin tai useat seuraavista ehdoista täyttyvät:

- Mitä-, miten- ja miksi-kysymykset ovat keskeisiä.
- Tutkijalla on vain vähän kontrollia tapahtumiin.
- Aiheesta on tehty vähän empiiristä tutkimusta.
- Tutkimuskohteena on jokin tämän ajan elävässä elämässä oleva ilmiö.

Ekstensiivisessä tapaustutkimuksessa pyritään löytämään ilmiöitä tai prosesseja koskevia yhteisiä ominaisuuksia ja yleisiä malleja. Päähuomio kiinnitetään teoreettisiin käsitteisiin, ei niinkään tosielämän ja tapaustoimijoiden omien näkökulmien ymmärtämiseen. (Eriksson & Koistinen, 2014.)

Tutkimuksessa kartoitetaan yritysten saatavilla oleva julkisen sektorin ja yksityisten yritysten tarjoama avoin data ja maksullinen data. Kartoituksella luodaan tutkimushetkeä vastaava ajantasainen kuva saatavilla olevasta datasta. Tutkimuksen tuloksia tarkastellaan case-yrityksen näkökulmasta. Tutkimuksen tapaus on laajennettavissa koskemaan data-analytiikkaa hyödyntäviä yrityksiä yleisesti ja tutkimuksessa esitettävät tulokset pätevät myös tapauksena toimivan organisaation ulkopuolella.

Avoimen datan ja maksullisen datan lähteitä etsitään Suomen kansallisesta -ja alueellisista portaaleista sekä julkishallinnon organisaatioiden ja yksityisten yritysten kotisivuilta. Tutkimuksessa kerätty datan tietoaineisto muodostaa tutkimuksen määrällisen aineiston.

## 5.1 Aineisto

Tutkimuksessa kerättävät tietoaineistot kootaan Excel-tiedostoon, mikä mahdollistaa työn tavoitteena olevan aineiston koneellisen jatkokäytön. Tarjolla olevasta datasta poimittavat tiedot ovat aineiston tarjoaja, tietokanta tai palvelu, josta aineisto tarjotaan, aineiston kuvailutieto, tieto aineiston avoimuudesta tai maksullisuudesta, tieto onko aineisto ohjelmallisesti hyödynnettävissä, tieto onko aineistossa rajapinta tai tarjotaanko aineisto tiedostolatauksena, aineiston tiedostotyyppi, aineiston kategoria sekä linkki aineistoon. Yksittäisestä rivistä, johon on täytetty aineistoa koskevat tiedot, käytetään tutkimuksessa termiä aineistorivi.

Aineiston ohjelmallisesti hyödynnettävyyttä tarkastellaan avoimen datan kriteerien näkökulmasta. Datan koneelliseen tarkasteluun ja hyödyntämiseen sopivat esimerkiksi CSV-, XLS- tai XML-muodot sekä erilaiset rajapinnat suoraan datalähteeseen, kun taas PDF-dokumenteissa tai HTML-sivuilla olevaa tietoa on vaikea lukea ohjelmallisesti (HRI, 2017b). Aineistoa ei merkitä koneellisesti hyödynnettäväksi, mikäli aineisto on tarjolla PDF-dokumenttina tai HTML-sivuna.

Mikäli aineisto on ohjelmallisesti hyödynnettävissä, lisätään tieto, tarjotaanko aineisto tiedostolatauksena vai ohjelmointirajapinnan kautta. Vastaavasti lisätään tieto tarjolla olevan aineiston tiedostotyyppistä. Mikäli aineisto ei ole ohjelmallisesti hyödynnettävissä, ei täytetä tietoja tiedostolatauksesta tai ohjelmointirajapinnasta eikä tiedostotyyppistä. Mikäli aineisto sisältää useita tarjolla olevia tiedostotyyppisiä, tehdään jokaisesta tiedostotyyppistä oma aineistorivi.



Aineiston kategorialla luokitellaan aineisto teemoihin, joita ovat esimerkiksi energia, liikenne ja terveys. Kategorioiden perusteella aineistoon voidaan tehdä hakuja ja todeta esimerkiksi millaista dataa on tarjolla terveyteen liittyen. Aineiston kategorisoinnissa käytetään Suomen kansallisen portaalin, Avoindata.fi-palvelun mukaisia kategorioita. Avoindata.fi-palvelu käyttää puolestaan hieman mukautettua versiota Euroopan data-portaalin (EDP) avoimen datan tietoaineistoille määrittelemistä kategorioista (HRI, 2017a). Palveluun on lisätty muutama ylimääräinen kategoria helpottamaan datan löytämistä. Samankaltaisten kategorioiden ansiosta Avoindata.fi-palveluun ladatut tietoa-ineistot määrittyvät oikeisiin kategorioihin myös EDP:n sivuilla. Lisäksi tietoa-ineistoja on helppo löytää myös muiden maiden portaaleista, kun kaikkialla Euroopassa pyritään käyttämään samantapaisia kategorioita. (Avoindata.fi, n.d.) Tutkimuksessa käytettävillä kategorioilla tuetaan puolestaan yhtenäistä käytäntö Suomessa. Tutkimuksessa käytetyt kategoriat esitetään taulukossa 1. Mikäli aineisto koskee useaa kategoriaa, tehdään jokaisesta kategoriasta oma aineistorivi.

**Taulukko 1.** Tietoa-ineistoissa käytetyt kategoriat (mukaillen Avoindata.fi, n.d.).

Kategoria	Kuvaus
Alueet ja kaupungit	Kartat, alueellinen data (data, joka kattaa tietyn maantieteellisen alueen)
Energia	Energiapolitiikka, energiateollisuus, energiankulutus, uusiutuva energia
Hallinto ja julkinen sektori	Vaalit ja äänestäminen, politiikka, hallinto ja päätöksenteko, vieraanvaraisuus ja lahjat, julkisen sektorin data, sanastot
Kansainväliset alueet	Ulkopolitiikka, kansainväliset järjestöt ja sopimukset, konfliktit ja rauha, asevoimat
Koulutus ja urheilu	Opetus ja koulutus, varhaiskasvatus, urheilu
Kulttuuri, taide ja vapaa-aika	Kirjastot, galleriat, museot, musiikki, arkistot, teatteri, harrastukset ja vapaa-aika
Liikenne	Liikennepolitiikka, liikennevälineet ja -muodot, julkinen liikenne, kävely ja pyöräily

Maatalous, kalastus, metsätalous ja elintarvikkeet	Maatalous, metsätalous, kalastus ja kaikki niihin liittyvä politiikka, sääntely, ohjeistukset ja toiminta, elintarvikkeet ja elintarviketeknologia
Matkailu ja turismi	Matkailukohteet, nähtävyydet, matkailutilat
Oikeus, oikeusjärjestelmä ja yleinen turvallisuus	Lakiasiat, lainopilliset ammatit, oikeusjärjestelmä, yleinen turvallisuus
Rakennettu ympäristö ja infrastruktuuri	Rakennukset ja rakentaminen, tiet, rautatiet, liikenneväylät, sähköverkot, puistot, muu infrastruktuuri
Talous ja rahoitus	Talouspolitiikka, verotus, rahoitus ja investoinnit, kulutus, kauppa ja kauppapolitiikka, tariffit, elinkeinot ja yritykset
Terveys	Lääketiede, sairaustiedot, ravitsemus, terveydenhuollon ammatit
Tiede ja teknologia	Tutkimukset (myös kysely- ja haastattelututkimukset), tutkimustieto ja -tulokset, teknologia
Väestö ja yhteiskunta	Väestön koostumus, työllisyys, sosiaalipolitiikka, asuminen, maahanmuutto, sosiaaliturva
Ympäristö ja luonto	Sää, luonnonympäristö, ympäristön pilaantuminen, jäte- ja vesihuolto

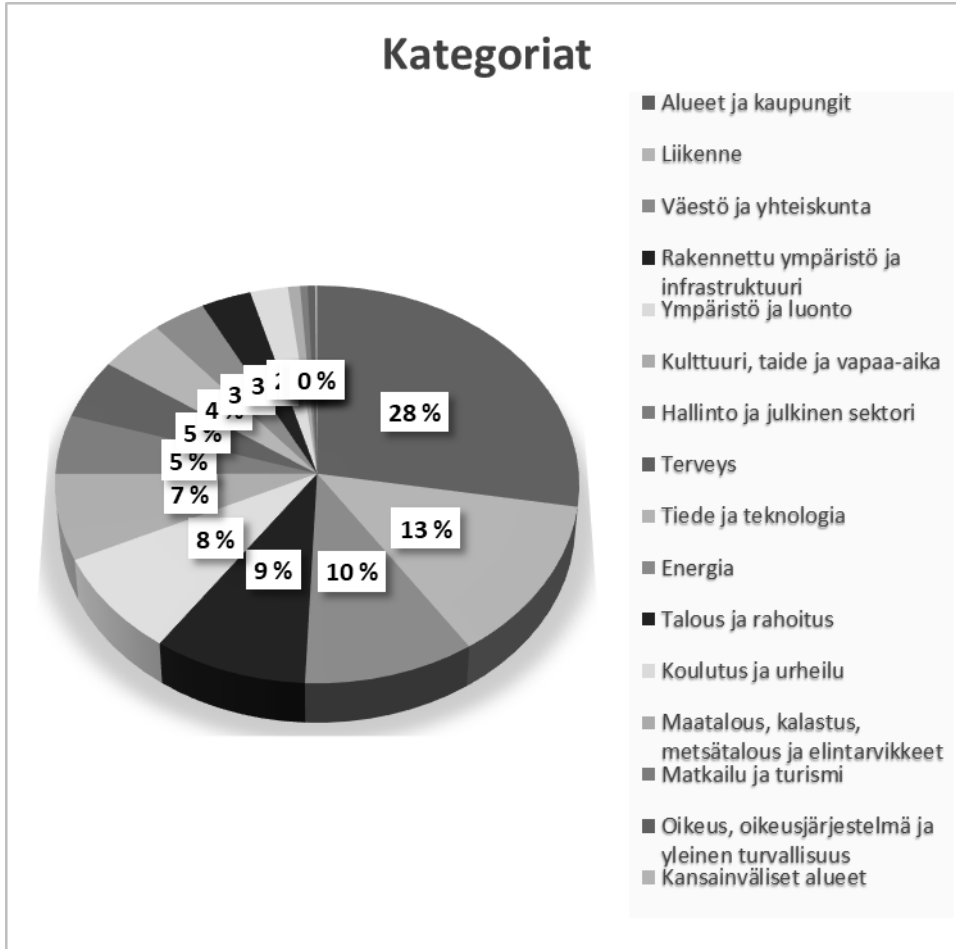
## 6 Tulokset

Tutkimukselle asetetut tutkimuskysymykset olivat:

- o Mitä avointa dataa ja maksullista dataa on tarjolla?
- o Onko tarjolla oleva data ohjelmallisesti hyödynnettävissä?
- o Mikä on yritysten esteenä tarjota avointa dataa ja maksullista dataa?

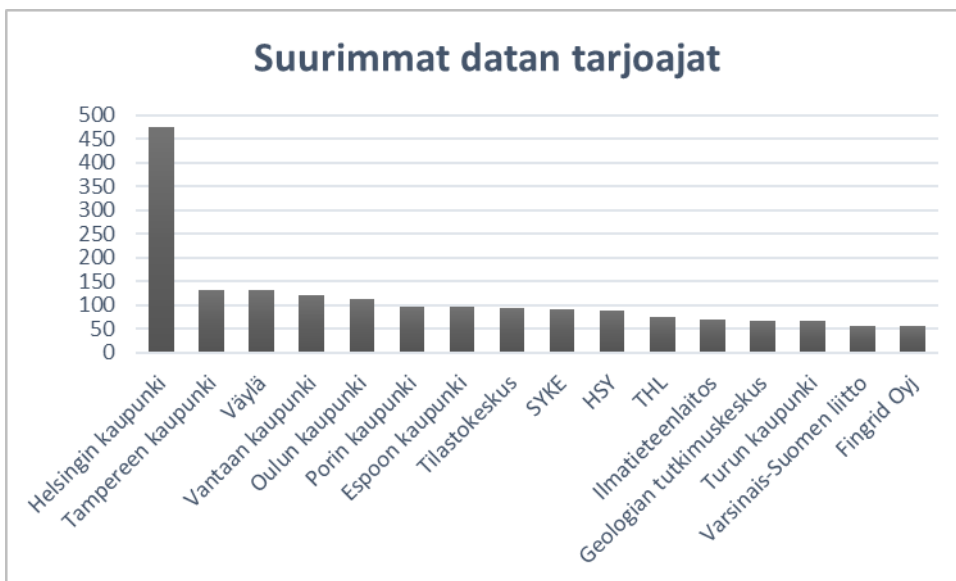
Tarjolla olevaan avointa dataa ja maksullista dataa analysoitiin tietoaaineistosta, joka kerättiin käymällä läpi avoimen datan jakelukanavat, Suomen kansallisen tason portaali, Avoindata.fi, sekä alueelliset dataportaalit. Tietoaaineistoa etsittiin myös datan avoimuuden edistäjien -sekä yritysten ja organisaatioiden verkkosivuilta ja palveluista. Tutkimuksessa kerättiin yhteensä 2278 aineistoriviä 125 tarjoajalta.

Analyysin tuloksena Avoimen datan osuus kerätyistä aineistoriveistä oli 96 %. Vastaavasti maksullisen datan osuus oli 4 %. Suurimmat avoimen datan ja maksullisen datan kategoriat on esitetty kuviossa 9. Kategorioista alueet ja kaupungit, väestö ja yhteiskunta sekä liikenne kattoivat 51 % koko aineistosta. Kategorioista selvästi suurin oli alueet ja kaupungit, kattaen lähes kolmasosan (28 %) koko aineistosta. Pienimpiä kategorioita olivat kansainväliset alueet (0,13 %), oikeus, oikeusjärjestelmä ja yleinen turvallisuus (0,48 %) sekä matkailu ja turismi (0,48 %).



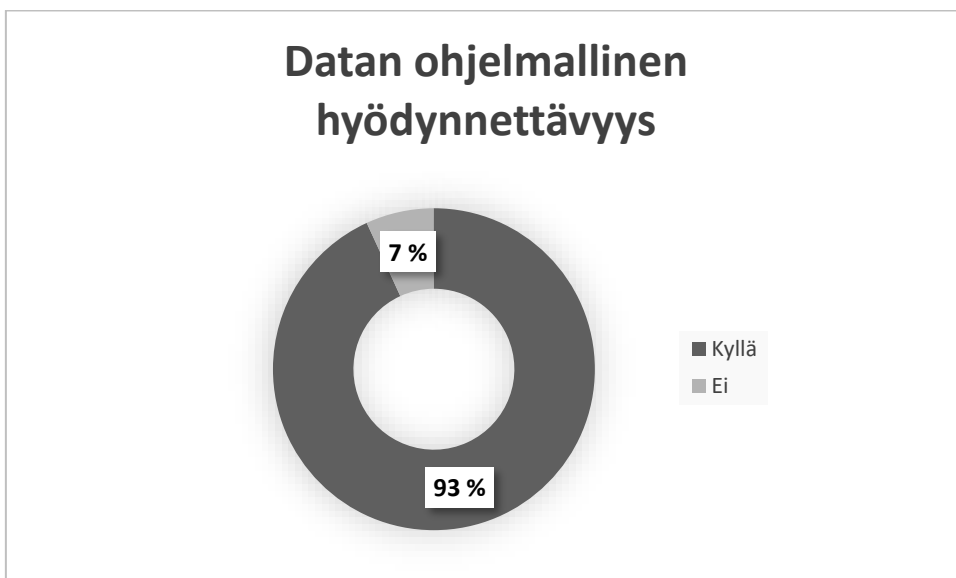
**Kuvio 9.** Aineisto kategorioittain.

Julkishallintoon kuuluvien organisaatioiden ja yritysten osuus kaikista datan tarjoajista oli 84 %. Vastaavasti julkishallintoon kuulumattomien yritysten ja organisaatioiden osuus oli 16 %. Suurimmat datan tarjoajat on esitelty kuviossa 10.



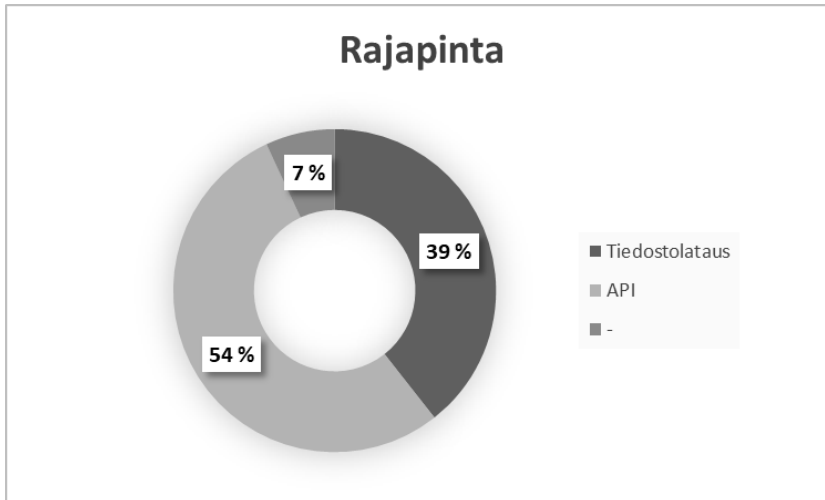
**Kuvio 10.** Suurimmat datan tarjoajat.

Analyysin perusteella tarjolla olevasta datasta on ohjelmallisesti hyödynnettävissä 93 % (kuvio 11). 7 % datasta ei ollut ohjelmallisesti hyödynnettävissä.



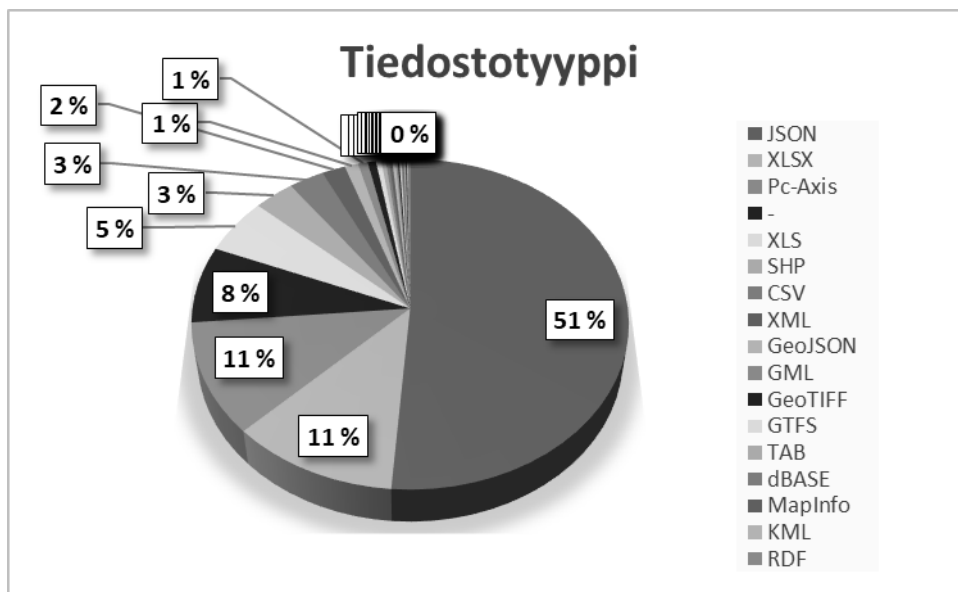
**Kuvio 11.** Datan ohjelmallinen hyödynnettävyys.

Tarjolla oleva data tarjotaan suurimmaksi osaksi tiedostolatauksena (kuvio 12). Tiedostolatauksen osuus oli 54 %. Datasta 39 % tarjotaan ohjelmointirajapinnan kautta ja 7 % datasta ei ollut ohjelmallisesti hyödynnettävissä.



**Kuvio 12.** Käytettävissä oleva rajapinta.

Aineiston tiedostotyypeistä suurimman osuuden muodostavat JSON, Excel sekä Pc-Axis. Yhdessä nämä muodostavat 78 % koko aineistosta. 8 % aineistoon ei merkattu tiedostotyyppiä, koska aineisto ei ole ohjelmallisesti hyödynnettävissä. Tiedostotyyppien osuudet on esitetty kuviossa 13.



**Kuvio 13.** Aineiston tiedostotyypit.

Kolmantena tutkimuskysymyksenä oli selvittää mikä on yritysten esteenä tarjota avointa dataa ja maksullista dataa. Tutkimuskysymykseen haettiin vastausta avoimeen dataan ja maksulliseen dataan liittyvästä kirjallisuudesta sekä tutkimuksista.

Kirjallisuudessa on esitetty syitä miksi yksityiset toimijat eivät ole halukkaista tarjoamaan avointa tai maksullista dataa. Ensinnäkin datan avaamiseen liittyy useita moniulotteisia lainsäädännöllisiä kysymyksiä, joista datan tarjoajan tulee olla tietoinen. Esimerkiksi tietovarantoihin ja perusrekistereihin liittyviä tai, niitä sivuavia lakeja on vähintäänkin kymmeniä. (Koski ja muut, 2017.) Euroopan komission selvityksen mukaan muita syitä yksityisten toimijoiden datan jakamattomuuteen ovat taloudellisten kannustimien puuttuminen, pelko kilpailuedun menettämisestä, epäluottamus siitä, että dataa käytetään sopimusehtojen mukaisesti, epätasapaino yritysten neuvotteluvoimissa, ja pelko siitä, että kolmannet osapuolet väärinkäyttävät tietoja (European Commission, 2020).

Herala (2018) esittää väitöskirjassaan väitteen, että yritykset eivät ole valmiita siirtämään datan avaamiseen. Syyksi Herala kertoo sen, että yritykset haluavat kontrolloida datan avulla luotujen tuotteiden innovointia sekä kaupallistamista. Datan avaaminen ei

tällöin ole yrityksille houkutteleva konsepti, johtuen kontrollin luovuttamisesta innovaatioprosessissa, epävarmat lopputulemat sekä hankalasti arvioidut tuotot ja muut hyödyt. Verhoevenin (2017) mukaan yksi este datan tarjoamiseen on yritysten teknisen kyvykkyyden puute. Verhoevenin mukaan yritykset saattavat kohdata tiedonhallinnan haasteita, jotka vaativat usein räätälöityjä ratkaisuja.

## 6.1 Case-tutkimus

Case-tutkimuksessa tarkastellaan, miten tutkimuksessa analysoitua tietoaineistoa voidaan hyödyntää tapaukseksi valitussa organisaatiossa. Organisaationa toimii sisäministeriön ohjauksessa toimiva pelastustoimi. Case-tutkimus on hypoteettinen, eikä perustu olemassa olevaan tarpeeseen.

Tietoaineiston hyödynnettävyyttä tarkastellaan onnettomuuksien ennalta ehkäisyyn liittyvän analytiikan näkökulmasta. Analytiikan perusteella pelastustoimi voisi tehdä päätöksiä sekä päätösten mukaisia toimia. Case-tutkimuksessa ei kuitenkaan toteuteta itse analytiikkaa vaan tutkimuksessa keskitytään siihen, millaisia datalähteitä analysoidusta tietoaineistosta voidaan löytää.

Ennaltaehkäisevässä analytiikassa tarkastellaan seuraavia osa-alueita:

- Alueellisen rikollisuuden kehitys.
- Ympäristökatastrofeihin varautuminen.
- Ihmismassojen hallinta.

### *Alueellisen rikollisuuden kehitys*

Alueellisen rikollisuuden kehityksen analytiikassa on tarkoitus seurata tietyn alueen rikollisuustilastojen kehitystä. Analytiikkaa varten tarvittaviksi ulkoisen datan lähteiksi voidaan ajatella kuuluvan alueen väestön kehitystä, alueen ikäluokkaa sekä alueen palveluja tapahtumatarjontaa kuvaava data. Rikollisuutta korreloivana palveluntarjontana voitaisiin pitää anniskeluravintoloita ja vastaavana tapahtumatarjontana alueella



järjestettäviä festivaaleja. Analytiikan ajatuksena on se, että valittuja tietoaaineistoja voidaan korreloida rikostilastojen kanssa ja saada havaintoja syy-seurausyhteyksistä väestön, palveluiden ja tapahtumatarjonnan kanssa. Analytiikan avulla voitaisiin tällöin kehittää alueen rikollisuuden ennusteita ja rikollisuuteen varautumista.

Alueellisen rikollisuuden kehityksen analytiikassa käytettävät ulkoisen datan lähteet haetaan tutkimuksessa analysoidusta tietoaaineistosta. Tietoaaineistosta löytyvät datan lähteet on esitetty kuvassa 1.

Kategoria	Lukumäärä
<input type="checkbox"/> <b>Kulttuuri, taide ja vapaa-aika</b>	<b>159</b>
<input type="checkbox"/> <b>Helsingin kaupunki</b>	<b>47</b>
<input type="checkbox"/> Helsinki Marketing	2
<input type="checkbox"/> <b>Helsingin kaupunki - tietoa matkailuun liittyvistä kohteista. Nähtävyydet, kohteet, majoitusliikkeet, juhla- ja kokoustilat, ravintolat, ostos- ja liikuntapaikat.</b>	<b>1</b>
<input type="checkbox"/> Avoin	1
<input type="checkbox"/> <b>Tiedostolataus</b>	<b>1</b>
XML	1
<input type="checkbox"/> <b>MyHelsinki Open API – paikat, tapahtumat ja aktiviteetit. MyHelsinki Open API on Helsinki Marketingin ylläpitämä avoin rajapinta, joka tarjoaa ajantasaista tietoa Helsingin sekä pääkaupunkiseudun paikoista, tapahtumista sekä aktiviteeteista.</b>	<b>1</b>
<input type="checkbox"/> Avoin	1
<input type="checkbox"/> <b>MyHelsinki Open API</b>	<b>1</b>
JSON	1
<input type="checkbox"/> <b>Kaupunginkanslia</b>	<b>3</b>
<input type="checkbox"/> <b>Helsingin tapahtumarajapinta.</b>	<b>1</b>
<input type="checkbox"/> Avoin	1
<input type="checkbox"/> <b>API</b>	<b>1</b>
JSON	1

[-] Väestö ja yhteiskunta	229
[-] HSY	3
[-] Ympäristöpalvelut	2
[-] Pääkaupunkiseudun väestötietoruudukko. Väestötietoruudukko sisältää ruutukohtaista tietoa	
[-] pääkaupunkiseudun väestön lukumäärästä, ikäjakaumasta sekä asumisväljyydestä.	1
[-] Avoin	1
[-] Tiedostolataus	1
SHP	1
[-] Tilastokeskus	15
[-] Paavo	3
Postinumeroalueittainen tilastoaineisto ja kartta-aineisto.	
Postinumeroalueittaiset paikkatietoaineistot: Tilastoaineisto yhdistettynä postinumeroalueittaiseen kartta-aineistoon, Postinumeroalueittainen kartta-aineisto, Merialueille ulottuva postinumeroalueittainen kartta-aineisto.	
Tilastoaineisto: Tietoja asukasrakenteesta, koulutusasteesta, asukkaiden ja talouksien tuloista, talouksien koosta ja	
[-] elämänvaiheesta, rakennuksista ja asunnoista, työpaikoista, asukkaiden pääasiallisesta toiminnasta.	3
[-] Avoin	3
[-] PxWeb API	1
JSON	1
[-] WFS API	1
JSON	1
[-] WMS API	1
JSON	1
[-] Ruututietokanta	3
Koordinaattipohjaisia tilastoaineistoja karttaruutuihin laskettuna: Asukasrakenne Koulutus Asuminen Tuloista ja työpaikoista.	
[-] Data on saatavissa koko Suomen alueelta sekä 250 m x 250 m-, 1 km x 1 km- tai 5 km x 5 km -	3
[-] Maksullinen	3
[-] Tiedostolataus	3
dBASE	1
MapInfo	1
SHP	1

**Kuva 1.** Kuva alueellisen rikollisuuden kehityksen analytiikassa käytettävistä ulkoisen datan lähteistä.

Alueen palveluntarjontaa koskeva data löytyy Kulttuuri, taide ja vapaa-aika -kategoriasta. Palveluntarjonta jakautuu alueellisiin datalähteisiin, joista taulukossa esimerkkinä Helsingin kaupungin tietoaaineisto. Väestöä koskeva data löytyy Väestö ja yhteiskunta -kategoriasta. Koko Suomen alueelta koskevaa tietoaaineistoa tarjoaa Tilastokeskus sekä postinumeroalueittain että ruututietokantana. HSY tarjoaa väestön ikäjakauman tietoaaineistoa pääkaupunkiseudun alueelta. Tietoaaineisto on tarjolla sekä avoimena datana että maksullisena datana. Aineisto on ohjelmallisesti hyödynnettävissä ohjelmistorajapintoina sekä tiedostolatauksina.

### Ympäristökatastrofeihin varautuminen

Ympäristökatastrofeihin varautumisen analytiikassa on tarkoitus seurata asutuskeskusten läheisyyttä teollisuusalueisiin sekä varautua mahdollisiin teollisuusalueilla syntyneisiin vaaratilanteisiin. Ympäristökatastrofeissa on tärkeää informoida ja turvata välittömässä vaikutuspiirissä olevat asukkaat. Analytiikkaa varten tarvittaviksi ulkoisen datan lähteiksi voidaan ajatella kuuluvan teollisuusalueiden ja asutuskeskusten sijaintitiedot sekä ilmanlaadun ja vesistöjen mittaustulokset. Analysoidusta tietoaaineistosta löytyvät datan lähteet on esitetty kuvassa 2.

Kategoria	Lukumäärä
<input type="checkbox"/> Alueet ja kaupungit	630
<input type="checkbox"/> SYKE	37
<input type="checkbox"/> Jätevedenpuhdistamot	1
Yhdyskuntajätevesidirektiivin mukaiset jätevedenpuhdistamot. Yhdyskuntajätevesidirektiivin mukaiset jätevedenpuhdistamot -aineisto sisältää Yhdyskuntajätevesidirektiivin (91/271/ETY) mukaiset taajamat,	1
<input type="checkbox"/> jätevedenpuhdistamot ja niiden purkupisteet.	1
<input type="checkbox"/> Avoin	1
<input type="checkbox"/> Tiedostolataus	1
-	1
<input type="checkbox"/> Teolliset toiminnot	1
EU-rekisteri teollisten toimintojen sijoittumisesta Suomessa. EU-rekisterin tarkoituksena on muodostaa	1
<input type="checkbox"/> ajantasainen rekisteri EU:n alueella sijaitsevista suurista teollisuuslaitoksista ja eläinsuojista.	1
<input type="checkbox"/> Avoin	1
<input type="checkbox"/> WMS API	1
XML	1
<input type="checkbox"/> Tilastokeskus	16
<input type="checkbox"/> Tuotanto- ja teollisuuslaitokset	1
<input type="checkbox"/> Tuotanto- ja teollisuuslaitokset sisältää Tilastokeskuksen yritysrekisteriin perustuvan toimipaikka-aineiston.	1
<input type="checkbox"/> Avoin	1
<input type="checkbox"/> WFS API	1
JSON	1

Ympäristö ja luonto	188
Ilmatieteenlaitos	25
Reaaliaikaiset havainnot	15
Ilman radioaktiivisuusvalvonta.	1
Avoin	1
WFS API	1
JSON	1
Kaupunkien ja teollisuuden ilmanlaatuhavainnot. Aineisto sisältää reaaliaikaiset asemakohtaiset	
ilmanlaatuhavainnot Suomen kaupungeista ja teollisuusalueilta (SO <sub>2</sub> , NO <sub>2</sub> , O <sub>3</sub> , TRS, CO, PM <sub>10</sub> , PM <sub>2.5</sub> ).	1
Avoin	1
WFS API	1
JSON	1
SYKE	35
Ilmansaasteiden päästökartat	1
Ilmansaasteiden ja kasvihuonekaasujen päästöt Suomessa vuosille 2015 ja 2030. Mukana olevia päästöjä	
ovat mm. hiukkaset, musta hiili, NO <sub>x</sub> , SO <sub>2</sub> , VOC, BaP, CO <sub>2</sub> ja CH <sub>4</sub> .	1
Avoin	1
-	1
-	1
Järvirajapinta	1
Järvirajapinta. Järvirajapinta sisältää tunniste-, fysiografia- ja säännöstelytietoja kaikista yli 1 ha kokoisista	
vedenkokoumista Suomessa. Tunnistetietoja ovat nimi, koordinaatit ja valuma-aluejaon mukainen tunnus.	
Fysiografiatietoja ovat esim. vesiala, rantaviivan pituus ja suurin syvyys. Säännöstelyistä kerrotaan mm.	
säännöstelyn tarkoitus ja säännöstelyluvan tietoja.	1
Avoin	1
OData API	1
JSON	1
Jätevedenpuhdistamot	1
Yhdyskuntajätevesidirektiivin mukaiset jätevedenpuhdistamot. Yhdyskuntajätevesidirektiivin mukaiset	
jätevedenpuhdistamot -aineisto sisältää Yhdyskuntajätevesidirektiivin (91/271/ETY) mukaiset taajamat,	
jätevedenpuhdistamot ja niiden purkupisteet.	1
Avoin	1
Tiedostolataus	1
-	1
Meriveden laatutietojen rajapinta	1
Meriveden laatutietojen rajapinta. Rajapinnan kautta julkaistaan fysikaalis-kemiallisia määrittelytuloksia	
Itämeren alueelta.	1
Avoin	1
OData API	1
JSON	1
Pintaveden laatutietojen rajapinta	1
Pintaveden laatutietojen rajapinta sisältää fysikaalis-kemiallisia	
määrittelytuloksia vedenlaatunäytteistä, joita on otettu Suomen järvistä, joista ja muista pienvesistä sekä	
Itämereltä.	1
Avoin	1
OData API	1
JSON	1

**Kuva 2.** Kuva ympäristökatastrofeihin varautumisen analytiikassa käytettävistä ulkoisen datan lähteistä.

Teollisuusalueiden sijaintitietoja koskeva data löytyy Alueet ja kaupungit -kategoriasta. Tietoaineistoa tarjoavat SYKE ja Tilastokeskus. Kaupunkikohtaisia tietoja löytyy myös kaupunkien omista tietoaineistoista. Ilmanlaadun ja vesistöjen mittaustuloksia koskeva data löytyy Ympäristö ja luonto -kategoriasta. Tietoaineistoa tarjoavat Ilmatieteenlaitos sekä SYKE. Asutuskeskuksia koskevan datan lähteinä voidaan pitää taulukossa 2 esitettyjä

tietoaineistoja. Tietoaineisto on tarjolla avoimena datana. Aineisto on ohjelmallisesti hyödynnettävissä ohjelmistorajapintoina sekä tiedostolatauksina.

### *Ihmismassojen hallinta*

Ihmismassojen hallinnan analytiikassa on tarkoitus seurata ja ennustaa ihmismassojen sijaintitietoja sekä varautua kokoontumisiin liittyvien vaaratilanteiden hallintaan. Analytiikan ajatuksena on se, että ihmismassojen kokoontumiset yleistyvät alueella järjestettävien tapahtumien yhteydessä. Myös vallitsevalla säällä voidaan ajatella olevan korrelaatiota kokoontumisiin. Analytiikkaa varten tarvittaviksi ulkoisen datan lähteiksi voidaan ajatella kuuluvan järjestettävät tapahtumat, säätiedot sekä paikannustiedot. Analysoidusta tietoaineistosta löytyvät datan lähteet on esitetty kuvassa 3.

Kategoria	Lukumäärä
<input checked="" type="checkbox"/> Alueet ja kaupungit	630
<input checked="" type="checkbox"/> Telia Company AB	4
<input checked="" type="checkbox"/> Crowd Insights	3
Ihmisvirta-analyysi kuvaa missä ja mihin aikaan väkijoukot liikkuvat. Ihmisvirta-analyysissä MTC-ruutu on pienimmillään 250 m x 250 m. Analyysi tuottaa tietoaineistoa ihmisvirran määrästä solun läpi tiettyinä ajanhetkenä. Data kerätään mobiililaitteiden anonyymeista sijaintitiedoista.	1
<input type="checkbox"/> Maksullinen	1
<input checked="" type="checkbox"/> API	1
-	1
Sijaintianalyysi kuvaa kävijöiden määrän tietyllä alueella ja tiettyinä ajanhetkenä. Data kerätään mobiililaitteiden anonyymeista sijaintitiedoista.	1
<input checked="" type="checkbox"/> Maksullinen	1
<input checked="" type="checkbox"/> API	1
-	1
<input checked="" type="checkbox"/> Tiede ja teknologia	98
<input checked="" type="checkbox"/> Ilmatieteenlaitos	27
<input checked="" type="checkbox"/> Ennustemallit	10
Sääennustedata RCR Hirlam. Aineisto sisältää Ilmatieteen laitoksen tuottaman tuoreimman sääennustedatan RCR Hirlam. Aineisto kattaa mm. maanpinnan läheiset perussääsuureet kuten 2m-lämpötilan, 10m-tuulen nopeuden ja suunnan, pintapaineen, pilvisyyden ja sademäärän Tuorein ennuste on 48 tuntia eteenpäin.	1
<input checked="" type="checkbox"/> Avoin	1
<input checked="" type="checkbox"/> WFS API	1
JSON	1
<input checked="" type="checkbox"/> Reaaliaikaiset havainnot	17
Säähavainnot. Aineisto sisältää Ilmatieteen laitoksen säähavaintoasemien hetkelliset pintasäähavaintoarvot. Pintasäähavaintoarvoja ovat lämpötila, paine, suhteellinen kosteus, tuulen nopeus ja suunta sekä sademäärä.	1
<input type="checkbox"/> Avoin	1
<input checked="" type="checkbox"/> WFS API	1
JSON	1

**Kuva 3.** Kuva ihmismassojen hallinnan analytiikassa käytettävistä ulkoisen datan lähteistä.

Paikannustietoja koskeva data Alueet ja kaupungit -kategoriasta. Paikannustietoja koskevaa tietoaaineisto tarjoaa Telia Company AB. Säättietoja koskeva data löytyy Tiede ja teknologia -kategoriasta. Tietoaaineistoa tarjoaa Ilmatieteenlaitos. Tapahtumatietoja koskeva data on esitetty taulukosta 2. Data löytyy Kulttuuri, taide ja vapaa-aika -kategoriasta. Tietoaaineistot jakautuvat alueellisiin datalähteisiin. Tietoaaineistoa tarjoavat ainakin Suomen suurimmat kaupungit. Paikannusta koskeva tietoaaineisto on tarjolla maksullisena datana ja säättietoja koskeva tietoaaineisto avoimena datana. Aineisto on ohjelmallisesti hyödynnettävissä ohjelmistorajapintojen kautta.

Tietoaaineistoon voi valita kerätyistä tiedoista näytettäväksi käyttötapauskohtaisesti tarpeelliset tiedot. Case-tutkimuksessa tarpeelliset tiedot olivat aineiston tuottajan ja kuvauksen lisäksi aineiston maksullisuus sekä käytettävissä olevat rajapinnat ja tiedostomuodot. Case-tutkimuksen perusteella analysoitua tietoaaineistoa voidaan hyödyntää saatavilla olevan ulkoisen datan selvitykseen, joten tutkimuksen tulokset ovat jatkohyödynnettävissä.

## 7 Diskussio

Tutkimuksessa kerätystä aineistosta suurin osa koski julkishallinnon tarjoamaa avointa dataa, mikä tukee aiempia tutkimuksia, joissa on todettu datan avaamisen olevan varsin pitkällä keskeisiä perustietovarantoja hallinnoivissa organisaatioissa, kun taas yksityisten yritysten hallussa olevan datan jakaminen ei ole kehittynyt riittävässä laajuudessa (Koski ja muut, 2017; European Commission, 2020). 125 suurimman datan tarjoajan joukosta ensimmäinen yksityisen alan yritys on sijalla 33.

Tutkimus osoittaa myös kaupunkien aktivoitumisen datan avaamisessa. 6Aika-hankkeessa mukana olleet Suomen kuusi suurinta kaupunkia ovat 16 suurimman datan tarjoajien joukossa. Tulokset osoittavat avatun aineiston tukevan yhteiskunnan digitalisointia aineiston ollessa suurimmilta osin ohjelmallisesti hyödynnettävissä sekä yli puolet aineistosta ollessa saatavilla ohjelmistorajapinnan kautta.

Liiketaloudellisesta näkökulmasta tulokset osoittavat datan avaamisen keskittyvän tässä vaiheessa Suomen sisämarkkina-alueeseen. Ulkomarkkinoita koskevan data sekä pääosin Suomeen suuntautuvan matkailun ja turismin data koostivat pienimmät osuudet kerätystä tietoaaineistosta.

Myös Euroopan komissio on tunnistanut puutteita tiettyjen alojen datan saatavuudessa ja tunnisteilla aloilla on annettu alakohtaista lainsäädäntöä datan saatavuudesta, todettujen markkinapuutteiden korjaamiseksi. Puutteet koskevat muun muassa ajoneuvoalaa, maksupalvelujen tarjoajia, älymittauksen tietoja, sähköverkkodataa ja älykkäitä liikennejärjestelmiä. (European Commission, 2020.)

Aiemmin Euroopan unioni on määritellyt viideksi tärkeimmäksi avoimen datan osa-alueeksi tilastot, paikka-, liikenne- ja yritystiedot sekä maapalloa koskevan havaintotiedon (Koski ja muut, 2017). Nämä aiemmin tärkeimmiksi määritellyt avoimen datan osa-alueet edustavat analysoidussa aineistossa suurimpia kategorioita, joten osa-alueiden

avaaminen on toteutunut Suomessa ja Suomen liiketoiminta-aluetta koskevassa tietoa-  
neistossa.

Aineistokeräyksen aikana tuli selväksi, että toteutetulle tutkimukselle oli tarvetta. Aiem-  
missä tutkimuksissa ei tullut esiin vastaavaa ulkoisen datan listaa, jossa olisi mukana sekä  
julkishallinnon että yksityisten yritysten tarjoama tietoaaineisto. Tarjolla oleva tietoa-  
aineisto on hajautunut useaan eri paikkaan; Suomen kansallisen tason portaaliin, alueelli-  
siin portaaleihin, datan avoimuutta edistävien organisaatioiden verkkosivuille sekä yksi-  
tyisyriyten kotisivuille, mikä hankaloittaa yrityksen liiketoiminnan kannalta tarpeellisen  
tietoaaineiston etsintää.

Tutkimuksen toimeksiantajayritys voi hyödyntämään tutkimuksen tuloksena syntyneitä  
tietoaaineistoja uusien analyyttikaliiketoimintamahdollisuuksien ideoinnissa sekä asiakas-  
projektiensa tukena. Tietoaaineiston perusteella toimeksiantajayritys pystyy arvioimaan  
tarjolla olevaa ulkoista dataa osana asiakkaalle toimitettavaa analyyttikaratkaisua. Tieto-  
aineistoa tarkasteltiin case-yrityksen näkökulmasta ja tarkastelun perusteella, toimeksi-  
antaja yritys voi käyttää aineistoa asiakastapauksissa tarkastelua vastaavalla tavalla. Tut-  
kimustulokset antavat myös yleisesti ulkoisen datan hyödyntäjille tietoa tarjolla olevasta  
tietoaaineistosta sekä tietoaaineiston hyödyntämismahdollisuuksista eri käyttökohteissa.

Tutkimuksen aikana havaittiin, että erityisesti maksullisen datan lähteitä on vaikea löytää.  
Havainto on samassa linjassa aiemmin tehtyihin havaintoihin. EU:n datastrategiaa kos-  
kevassa tiedonannossa Euroopan komissio toteaa, että tällä hetkellä dataa ei ole saata-  
villa riittäviä määriä innovatiiviseen uudelleenkäyttöön ja erityisesti yksityisten yritysten  
hallussa olevan datan jakaminen ei ole kehittynyt riittävässä laajuudessa (European  
Commission, 2020).

Koski ja muut (2017) toteavat, että myös avoimen datan jakelu on pirstaleista, jonka takia  
datan löydettävyyttä on hyödyntäjille haasteellista. Julkisen sektorin organisaatioiden  
avoimen datan tarjoajien selvityksessä lähes kaikilla organisaatioilla oli oma avoimen



datan jakelukanava tai useampia kanavia, jonka lisäksi tietoa jaettiin Avoindata.fi -portaalissa. Yksi organisaatioista jakoi avointa dataa vain omien sivujensa kautta ja yksi vain Avoindata.fi -palvelun kautta. (Koski ja muut, 2017.)

Ulkoisen datan etsintää helpottaisi se, mikäli kaikki tarjottava data olisi samassa paikassa. Alueellisten portaalien tarpeellisuutta pohtii myös Jokela (2015). Vaihtoehdoksi hän esittää avoimen datan jakeluun pelkkää kansallisen tason portaalialueita. Alueellisten portaalien käytettävyyttä tukee se, että portaalit on toteutettu 6Aika-hankeessa avoimella lähdekoodilla ja palvelu on helposti monistettavissa muiden alueiden tai kuntien käyttöön (Jokela, 2015).

Avoindata.fi-palvelun käyttämistä datan tarjoamiseen tukee se, että valmiin portaalin käyttäminen säästää kuluja, ja säästyneen budjetin voi suunnata esimerkiksi datan avaamiseen, oman portaalin kehittämisen sijaan. Avoindata.fi:n ideana on tarjota palvelu ilmaiseksi ministeriöiden, kuntien, yhdistysten, yritysten ja yksityishenkilöiden käyttöön. Avoindata.fi -palvelussa julkaistu data saa lisäksi kansainvälistä näkyvyyttä, sillä kaikki palvelussa oleva data siirretään Euroopan dataportaaliiin. (Avoindata.fi, n.d.) Toisaalta tuntuu, että Avoindata.fi-palvelu toimii tietyissä tilanteissa pelkkänä mainospaikkana, sillä tutkimuksessa tuli esiin aineistoja, joissa ei yritystietojen lisäksi ollut ollenkaan dataa tarjolla.

Maksulliselle datalle tulisi tuottaa avointa dataa vastaavia portaalialueita. Koski ja muut (2017) ehdottavat Avoindata.fi-palvelun käytön markkinointia julkisten organisaatioiden lisäksi yritysten ja kolmannen sektorin datan avaamiseen. Maksullista dataa ei portaalissa kuitenkaan voida tarjota, koska data harvestoidaan palvelusta Euroopan dataportaaliiin. Palvelua voitaisiin laajentaa erillisellä osuudella, joka rajattaisiin harvestoinnin ulkopuolelle. Laajennuksen yhteydessä myös palvelun nimeä tulisi tarkastella uudelleen.

Aineistokeräyksessä yllätti aineistorivien lopullinen lukumäärä. Suuri aineistorivien lukumäärä liittyi keräyksellä pyrittiin tarkkuuteen. Aineistosta pyrittiin poimimaan omiksi

riveikseen kaikki kategoriat, joihin aineisto liittyi. Vastaavasti kaikki tarjolla olevat rajapinnat ja tiedostomuodot pyrittiin poimimaan omiksi riveikseen.

Tutkimuksen edetessä, keräykseen käytettävän resurssimäärän hallitsemiseksi, poimittaviin tietoihin jouduttiin tekemään tiettyjä yleistyksiä. Aineistojen rajapinnat ja tiedostomuodot priorisoitiin ja aineistoriveiksi poimittiin tarjolla olevista vaihtoehdoista ohjelmallisen jatkokäsittelyn näkökulmasta käyttökelpoisimmat vaihtoehdot. Ohjelmallisen jatkokäsittelyn vaihtoehtojen priorisointi tukee digitalisaatiota sekä avoimen datan ko- neluettavuuden kriteeriä. Aineistosta poimittiin omiksi aineistoriveiksi kaikki kategoriat, joihin aineisto liittyi.

Tutkimuksessa tehdyillä yleistyksillä on vaikutusta tutkimuksen reliabiliteettiin, koska tutkimus ei ole toteutettavissa täysin samanlaisena. Toistettavuuteen vaikuttaa myös tutkimuksen kohdeaineistossa tapahtuvat muutokset. Tutkimus on kuitenkin validi tutkimuksen tarkoituksen näkökulmasta ja tutkimus antaa ajantasaisen kuvan ulkoisen datan saatavuudesta. Aineistoista koottu Excel sopii tarkoitettuun käyttötarkoitukseen.

Aineiston kategorisoinnissa käytettiin Suomen kansallisen portaalin, Avoindata.fi-palvelun mukaisia kategorioita, jolla tuettiin Suomen ja Euroopan yhtenäistä käytäntöä. Kategorisointi oli tiettyjen aineistojen kohdalla haastavaa liittyen kategorioiden suhteellisen pieneen lukumäärään ja tätä kautta yksittäisten kategorian laajaan kattavuuteen. Jossain tilanteissa esiintyi myös tarvetta uusille kategorioille. Euroopan dataportalissa aihealueet-kategorian rinnalle on nyt ryhmät-kategoria, joka tarjoaa käyttäjien tarpeista nousseen vaihtoehtoisen tavan selata datajoukkoja.

Jatkotutkimuksena ehdotetaan kerätyn tietoaineiston ryhmittelyä kategorialla tarkemmalle tasolle helpottamaan aineiston jatkokäyttöä. Kerättyä tietoaineistoa voisi myös laajentaa muilla tarjottavilla rajapinnoilla ja tiedostomuodoilla. Tarjolla oleva ulkoinen data muuttuu nopeasti, jolloin kerättyä listaa tulisi päivittää aktiivisesti. Aineisto olisi

myös hyvä siirtää erilliseen tietokantaan ja aineiston käyttöön tulisi harkita käyttöliittymän rakentamista.

EU:n datan avaamisen edistäminen on suunnitelmallista ja Euroopan komissio esittää 19.2.2020 annetussa tiedonannossa datan avoimuutta edistävän datastrategian lisäksi miten komissio aikoo tukea ja edistää tekoälyn kehittämistä ja käyttöönottoa Euroopassa. Komission mukaan datan saatavuus on edellytys tekoälyjärjestelmien kouluttamiseksi, sillä tuotteet ja palvelut ovat nopeasti siirtymässä hahmontunnistuksesta ja päätelmien generoinnista pitemmälle vietyihin ennustemenetelmiin ja sitä kautta parempiin päätöksiin. (European Commission, 2020).

Rajakannas (2018b) esittää ajatuksen, jonka mukaan datan arvo voi lähitulevaisuudessa laskea. Arvon lasku johtuu Rajakannaksen mukaan kahdesta syystä. Ensimmäkin koneoppimisen alalla orastava läpimurto osittain ohjatusta oppimisesta (semi-supervised learning) mahdollistaa analyysien teon kohtuullisella määrällä dataa, valtaviin tietomääriin sijasta. Toinen muutokseen vaikuttava asia on kasvava paine julkishallinnon datan avaamiseen ja toisaalta paine tarjota ihmisille hallinta heidän itsensä muodostamaan dataan. Rajakankaan mukaan arvo on siirtymässä datasta datan perusteella tehtävien analyysien tuottamaan arvoon.

## Lähteet

- 6Aika. (n.d.). *Älykaupungit tehdään yhdessä*. Noudettu 2020-3-29 osoitteesta <https://6aika.fi/>
- Aho, M. (2011). *Konstruktio suorituskyvyn johtamisen kypsyyden arviointiin*. Tampereen teknillinen yliopisto. <http://urn.fi/URN:ISBN:978-952-15-2696-1>
- Akerkar, R. & Sajja, P. S. (2016). *Intelligent Techniques for Data Science*. Springer. <https://doi.org/10.1007/978-3-319-29206-9>
- Alma Talent. (2020). *Alma Talent Tietopalvelut*. Alma Media Oyj. Noudettu 2020-5-2 osoitteesta <https://www.almatalent.fi/tietopalvelut>
- Asiakastieto. (n.d.). *Luottotiedot ja -luokitukset*. Asiakastieto Group Oy. Noudettu 2020-4-13 osoitteesta <https://www.asiakastieto.fi/web/fi/palvelut-yrityksille/tarkista-ennen-kaupantekoa/luottotiedot-ja-luokitukset.html>
- Auvinen, E. (2020). *Kännyköidemme sijaintitiedot näyttävät ensi kertaa, miten korona pysäytti Tampereen ja miten elämä alkoi taas palata*. Aamulehti. Noudettu 2020-17-6 osoitteesta <https://www.aamulehti.fi/a/717c6107-54e9-4d8b-874c-6b3f6031a033?c=1522737894164>
- AVAA. (n.d.). *Tietoa palvelusta*. Opetus- ja kulttuuriministeriö. Noudettu 2020-4-26 osoitteesta <https://avaa.tdata.fi/web/avaa/tietoa-palvelusta>
- Avoindata.fi. (n.d.). *Suomen kaikki avoin data yhdestä paikasta*. Noudettu 2019-12-22 osoitteesta <https://www.avoindata.fi/fi>
- Barker, I. (2016). *Dark Data and Why You Should Worry About It*. BetaNews. Noudettu 2020-4-12 osoitteesta <https://betanews.com/2016/02/19/big-dark-data/>

- Becker, J., Knackstedt, R. & Pöppelbuß, J. (2009). Developing Maturity Models for IT Management. *Business & Information Systems Engineering*, 1(3), 213-222.  
<https://doi.org/10.1007/s12599-009-0044-5>
- Belissent, J. (2017). *The Age of Alt: Data Commercialization Brings Alternative Data To Market*. Forrester. Noudettu 2020-3-21 osoitteesta  
[https://go.forrester.com/blogs/17-06-22-the\\_age\\_of\\_alt\\_data\\_commercialization\\_brings\\_alternative\\_data\\_to\\_market/](https://go.forrester.com/blogs/17-06-22-the_age_of_alt_data_commercialization_brings_alternative_data_to_market/)
- Bisnode, (n.d.). *Ratkaisuja liiketoimintasi haasteisiin. Smart data auttaa sinua tekemään parempia päätöksiä*. Bisnode Finland Oy. Noudettu 2020-4-13 osoitteesta  
<https://finland.bisnode.fi/ratkaisut/>
- Blank, M. (2019). *Open Data Maturity Report 2019*. European Commission.  
<https://doi.org/10.2830/073835>
- Brownlow, J., Zaki, M., Neely, A. & Urmetzer, F. (2015). *Data and Analytics - Data-Driven Business Models: A Blueprint for Innovation*. University of Cambridge. Noudettu 2020-2-24 osoitteesta  
<https://cambridgeservicealliance.eng.cam.ac.uk/resources/Downloads/Monthly%20Papers/2015FebruaryPaperTheDDBMIInnovationBlueprint.pdf>
- Business Finland. (n.d.). Tutkimukset ja tilastot. Noudettu 2020-5-24 osoitteesta  
<https://www.businessfinland.fi/suomalaisille-asiakkaille/palvelut/matkailun-edistaminen/tutkimukset-ja-tilastot/tutkimukset-ja-tilastot-lyhyesti/>
- Business Tampere. (n.d.). *ITS Factory. The Future of Transport*. Noudettu 2020-4-5 osoitteesta  
<https://businesstampere.com/business-environment/business-ecosystems/mobility/>

- Carrara, W., Wae, S. C., Fischer, S. & van Steenberg, E. (2015). *Creating Value through Open Data: Study on the Impact of Re-use of Public Data Resources*. European Commission. <https://doi.org/10.2759/328101>
- CITO Research. (2012). *Mission: How to Find and Use External Data*. Noudettu 2020-3-22 osoitteesta <https://citoresearch.com/data-science/mission-how-find-and-use-external-data>
- CMMI. (2002). *Capability Maturity Model Integration (CMMI), Version 1.1*. Noudettu 2020-3-8 osoitteesta [https://resources.sei.cmu.edu/asset\\_files/TechnicalReport/2002\\_005\\_001\\_14072.pdf](https://resources.sei.cmu.edu/asset_files/TechnicalReport/2002_005_001_14072.pdf)
- Comuzzi, M. & Patel, A. (2016). How Organisations Leverage Big Data: A Maturity Model. *Industrial Management & Data Systems*, 116(8), 1468-1492. <https://doi.org/10.1108/IMDS-12-2015-0495>
- conzoom®. (n.d.). *conzoom®license*. Geomatic A/S. Noudettu 2020-4-26 osoitteesta <https://conzoom.com/license/>
- Creative Commons. (2019). *About CC Licenses*. Noudettu 2020-7-29 osoitteesta <https://creativecommons.org/about/ccllicenses/>
- Crossman, A. (2019). *Pros and Cons of Secondary Data Analysis. A Review of the Advantages and Disadvantages in Social Science Research*. ThoughtCo. Noudettu 2020-3-22 osoitteesta <https://www.thoughtco.com/secondary-data-analysis-3026536>
- DataBusiness.fi. (n.d.) *Avoin data*. Noudettu 2020-4-11 osoitteesta <https://www.databusiness.fi/fi/avoindata/>

- Davis, J. (2019). *How to Buy External Data to Fuel Analytics, AI Insights*. InformationWeek. Noudettu 2020-3-24 osoitteesta <https://www.informationweek.com/strategic-cio/it-strategy/how-to-buy-external-data-to-fuel-analytics-ai-insights/d/d-id/1334023>
- Day Thomson, S. (2016) *Preserving Social Media. DPC Technology Watch Report 16-01 February 2016*. Digital Preservation Coalition. <http://dx.doi.org/10.7207/twr16-01>
- Ellram, L. M. & Tate, W. L. (2016). The Use of Secondary Data in Purchasing and Supply Management (P/SM) Research. *Journal of Purchasing & Supply Management*, 22(4), 250-254. <http://dx.doi.org/10.1016/j.pursup.2016.08.005>
- Eriksson, P. & Koistinen, K. (2014). Monenlainen tapaustutkimus. *Kuluttajatutkimuskeskuksen tutkimuksia ja selvityksiä*, 11. <http://hdl.handle.net/10138/153032>
- EU:n avoimen datan portaali. (n.d.) Noudettu 2020-3-28 osoitteesta <https://data.europa.eu/euodp/fi/home>
- European Commission. (2020) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European Strategy for Data. Noudettu 2020-4-11 osoitteesta <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN>
- European Data Portal. (2020) Noudettu 2020-4-2 osoitteesta <https://www.europeandataportal.eu/en>

- Facebook. (n.d.) *Facebook Analytics*. Noudettu 2020-4-12 osoitteesta <https://www.facebook.com/help/analytics/1710582659188030>
- Fonecta. (n.d.). *Palvelut. Analytiikka*. Noudettu 2020-4-13 osoitteesta <https://www.fonecta.fi/palvelut/analytiikka>
- Forrester. (2017). *Digital Is Driving The Next Generation Of Data Marketplaces*. Noudettu 2020-3-21 osoitteesta <https://www.pitneybowes.com/content/dam/pitneybowes/us/en/white-papers/pitney-bowes-forrester-data-us.pdf>
- Forsström, P-L. (2019). *Uudistunut PSI-direktiivi tuo uutta puhtia saatavuuteen*. Avointiede. Noudettu 2020-3-26 osoitteesta <https://avointiede.fi/fi/ajankohtaista/uudistunut-psi-direktiivi-tuo-uutta-puhtia-saatavuuteen>
- Fält, K. (2016). *Someaineistojen arkistointi ja jatkokäyttö kaatuvat useimpien alustojen käyttöehtoihin*. Tietoarkisto. Noudettu 2020-4-13 osoitteesta [https://www.fsd.tuni.fi/lehti/fi/45/some\\_data.html](https://www.fsd.tuni.fi/lehti/fi/45/some_data.html)
- Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144. <http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007>
- Gartner. (n.d.). *Big Data*. Noudettu 2020-2-28 osoitteesta <https://www.gartner.com/en/information-technology/glossary/big-data>
- Gartner. (2017). *Leadership Vision for 2018: Data and Analytics Leader*. Noudettu 2020-3-12 osoitteesta <https://www.scribd.com/document/432918125/Gartner-Leadership-Vision-for-2018-Data-and-Analytics-Leaders>



- Gartner. (2018). *Gartner Survey Shows Organizations Are Slow to Advance in Data and Analytics*. Noudettu 2020-3-21 osoitteesta <https://www.gartner.com/en/newsroom/press-releases/2018-02-05-gartner-survey-shows-organizations-are-slow-to-advance-in-data-and-analytics>
- George, G., Haas, M. R. & Pentland, A. (2014). Big Data and Management. *Academy of Management Journal*, 57(2), 321-326. <http://dx.doi.org/10.5465/amj.2014.4002>
- Google. (2020). *Trendit ohjeet*. Noudettu 2020-4-15 osoitteesta [https://support.google.com/trends/answer/4365533?hl=fi&ref\\_topic=6248052](https://support.google.com/trends/answer/4365533?hl=fi&ref_topic=6248052)
- GTK. (n.d.). *Rajapintapalvelut*. Geologian tutkimuskeskus. Noudettu 2020-4-4 osoitteesta <https://www.gtk.fi/palvelut/aineistot-ja-verkkopalvelut/rajapintapalvelut/>
- Halper, F. & Krishnan, K. (2013). *TDWI Big Data Maturity Model Guide. Interpreting Your Assessment Score*. TDWI. Noudettu 2020-3-14 osoitteesta <https://tdwi.org/~media/3BF039A2F7E1464B8290D8A9880FEC22.pdfma>
- Halper, F. & Stodder, D. (2014). *TDWI Analytics Maturity Model Guide. Interpreting Your Assessment Score*. TDWI. Noudettu 2020-3-11 osoitteesta [https://www.microstrategy.com/getmedia/9b914607-084f-4869-ae64-e0b3f9e003de/TDWI\\_Analytics-Maturity-Guide\\_2014-2015.pdf](https://www.microstrategy.com/getmedia/9b914607-084f-4869-ae64-e0b3f9e003de/TDWI_Analytics-Maturity-Guide_2014-2015.pdf)
- Hartmann, P. M., Zaki, M., Fieldmann, N. & Neely, A. (2016). Capturing Value from Big Data – A Taxonomy of Data-Driven Business Models Used by Start-Up Firms. *International Journal of Operations and Production Management*, 36(10), 1382-1406. <https://doi.org/10.1108/IJOPM-02-2014-0098>

Helen Oy. (2020). *Avoin data*. Noudettu 2020-4-26 osoitteesta <https://www.helen.fi/helen-oy/vastuullisuus/ajankohtaista/avoindata>

Herala, A. (2018). *Benefits from Open Data: Barriers of Supply and Demand of Open Data in Private Organizations*. Lappeenrannan teknillinen yliopisto. <http://urn.fi/URN:ISBN:978-952-335-262-9>

HRI. (2017a). *Avoimen datan palvelu. Pääkaupunkiseudun julkinen tieto hyötykäyttöön*. Helsinki Region Infoshare. Noudettu 2020-3-18 osoitteesta <https://hri.fi/fi/>

HRI. (2017b). *Mitä on avoin data?*. Helsinki Region Infoshare. Noudettu 2020-6-17 osoitteesta <https://hri.fi/fi/ohjeet/mita-on-avoin-data/>

HSL. (n.d.). *Avoin data*. Helsingin seudun liikenne. Noudettu 2020-5-10 osoitteesta <https://www.hsl.fi/avoindata>

HSY. (2019). *Hae avointa dataa*. Helsingin seudun ympäristöpalvelut. Noudettu 2020-5-10 osoitteesta <https://www.hsy.fi/fi/asiantuntijalle/avoindata/Sivut/default.aspx>

Ilmatieteen laitos. (2019). *Ilmatieteen laitoksen avoin data ja lähdekoodi*. Noudettu 2020-4-4 osoitteesta <https://www.ilmatieteenlaitos.fi/avoin-data>

Itkonen, J. (2015). *Kiihdyttääkö digitalisaatio talouskasvua?* Suomen Pankki. Noudettu 2020-2-22 osoitteesta [https://helda.helsinki.fi/bof/bitstream/handle/123456789/13920/eurotalous\\_blogi\\_27.pdf?sequen](https://helda.helsinki.fi/bof/bitstream/handle/123456789/13920/eurotalous_blogi_27.pdf?sequen)

Intrum. (2019). *Informaatio- ja analytiikkapalvelut*. Noudettu 2020-4-13 osoitteesta <https://www.intrum.fi/fi/ratkaisut-yrityksille/palvelut-yrityksille/luottotiedot/>

- ITS Factory. (2020). *ITS Factory Developer Wiki*. Noudettu 2020-4-5 osoitteesta [http://wiki.itsfactory.fi/index.php/ITS\\_Factory\\_Developer\\_Wiki](http://wiki.itsfactory.fi/index.php/ITS_Factory_Developer_Wiki)
- Jantunen, L. & Koskinen, K. (2019). *Tekoäly ja massiiviset dataaineistot muovaavat rahoitusmarkkinoita* Suomen Pankki. Noudettu 2020-2-22 osoitteesta <https://helda.helsinki.fi/bof/bitstream/handle/123456789/16508/eurotalous-analyysiartikkeli-27-11-2019.pdf?sequence=1>
- Johnston, M. P. (2014). Secondary Data Analysis: A Method of which the Time Has Come. *Qualitative and Quantitative Methods in Libraries*, 3(3), 619-626. ISSN 2241-1925. Noudettu 2020-3-19 osoitteesta <http://www.qqml-journal.net/index.php/qqml/article/view/169>
- Jokela, S. (2015). *Alueellisilla dataportaaleilla menossa feeding frenzy*. DataBusiness.fi. Noudettu 2020-4-10 osoitteesta <https://www.databusiness.fi/fi/blogi/alueellisilla-dataportaaleilla-menossa-feeding-frenzy/>
- Kauhanen-Simanainen, A. & Suurhasko, M. (2015). Avoimesta datasta innovatiiviseen tiedon hyödyntämiseen. Avoimen tiedon ohjelman 2013–2015 loppuraportti. *Valtiovarainministeriön julkaisuja*, 31/2015. Noudettu 2020-3-28 osoitteesta <https://vm.fi/documents/10623/1107406/Avoimen+tiedon+ohjelman+loppuraportti/>
- Koski, H., Honkanen, M., Luukkonen, J., Pajarinen, M. & Ropponen, T. (2017). Avoimen datan hyödyntäminen ja vaikuttavuus. *Selvitys- ja tutkimustoiminnan julkaisusarja*, 2017(40). Noudettu 2019-12-23 osoitteesta <https://vnk.fi/julkaisu?pubid=18703>

- Leem, C. S., Kim, B. W., Yu E. J. & Paek, M H. (2008). Information Technology Maturity Stages and Enterprise Benchmarking: and Empirical Study. *Industrial Management & Data Systems*, 108(9), 1200-1218. <https://doi.org/10.1108/02635570810914892>
- Leinonen, P. (2019). *Avoin data – avain innovaatioihin*. Fingrid. Noudettu 2020-4-11 osoitteesta <https://www.fingridlehti.fi/avoin-data-innovaatiot/>
- LIPASTO. (n.d.). LIPASTO - Suomen liikenteen pakokaasupäästöjen ja energiankulutuksen laskentajärjestelmä. Noudettu 2020-5-9 osoitteesta <http://lipasto.vtt.fi/index.htm>
- Lounaistieto.fi. (n.d.). *Lounaistieto. Avoimesti*. Noudettu 2020-4-4 osoitteesta <https://www.lounaistieto.fi/tietoa-meista/>
- Lowell. (2020). *Informaatiopalvelut*. Noudettu 2020-5-2 osoitteesta <https://www.lowell.fi/ratkaisut/informaatiopalvelut/>
- Luke. (n.d.). *Avoin tieto*. Luonnonvarakeskus. Noudettu 2020-4-5 osoitteesta <https://www.luke.fi/avoin-tieto/>
- Maanmittauslaitos. (n.d.). *Avoimien aineistojen tiedostopalvelu*. Noudettu 2020-4-4 osoitteesta <https://www.maanmittauslaitos.fi/asioi-verkossa/avoimien-aineistojen-tiedostopalvelu>
- Marjamäki, V-M. (2014). Tiedon hyödyntämisen ekosysteemi. *Liikenneviraston tutkimuksia ja selvityksiä*, 2014(56). <http://urn.fi/URN:ISBN978-952-317-041-4>

- Marr, B. (2018). *Here's Why Data Is Not The New Oil*. Forbes. Noudettu 2020-6-22 osoitteesta <https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/#312e32763aa9>
- Mettler T., Rohner P. & Winter R. (2010) Towards a Classification of Maturity Models in Information Systems. In: D'Atri A., De Marco M., Braccini A., Cabiddu F. (eds) *Management of the Interconnected World*. Physica-Verlag HD. [https://doi.org/10.1007/978-3-7908-2404-9\\_39](https://doi.org/10.1007/978-3-7908-2404-9_39)
- Miele, S. & Shockley, R. (2013). *Analytics: The real-world use of big data. How innovative enterprises in the midmarket extract value from uncertain data*. IBM Global Business Services Noudettu 2020-2-28 osoitteesta [https://www.informationweek.com/pdf\\_whitepapers/approved/1372892704\\_analytics\\_the\\_real\\_world\\_use\\_of\\_big\\_data.pdf](https://www.informationweek.com/pdf_whitepapers/approved/1372892704_analytics_the_real_world_use_of_big_data.pdf)
- Nysse. (2020). *Avoin data*. Tampereen seudun joukkoliikenne. Noudettu 2020-5-13 osoitteesta <https://joukkoliikenne.tampere.fi/ohjeita-ja-tietoa/digipalvelut/avoin-data.html>
- OKFI. (2019). *Ensimmäinen avoimen datan vuosikymmen Suomessa -tapahtumat 8.-10.10*. Open Knowledge Finland. Noudettu 2020-4-7 osoitteesta <https://www.okf.fi/fi/2019/06/12/ensimmainen-vuosikymmen/>
- OKFI. (n.d.). *Tietoja OKFIsta*. Open Knowledge Finland. Noudettu 2020-4-11 osoitteesta <https://www.okf.fi/fi/tietoja/>
- Open Definition. (n.d.). *Open Definition 2.1*. Open Knowledge. Noudettu 2020-3-18 osoitteesta <https://opendefinition.org/od/2.1/en/>

- Oulu. (n.d.). *Oulun kaupungin dataportaali. Avointa dataa Oulusta*. Oulun kaupunki. Noudettu 2020-4-4 osoitteesta <https://data.ouka.fi/fi/>
- Posti. (2018). *Mikä on Katudata-tilannepalvelu?* Posti Group Oyj. Noudettu 2020-4-16 osoitteesta <https://www.posti.fi/fi/yrityksille/vinkit-ja-caset/dataa-ja-markkinointia/katudata-usein-kysyttya>
- Posti. (n.d.). *Dataa ja markkinointia. Tuloksia myyntiin ja markkinointiin asiakasdatalla*. Posti Group Oyj. Noudettu 2020-4-13 osoitteesta <https://www.posti.fi/fi/yrityksille/dataa-ja-markkinointia>
- PRH. (n.d.). *Tietoa avoimen datan rajapinnoista*. Patentti- ja rekisterihallitus. Noudettu 2020-4-11 osoitteesta <http://avoindata.prh.fi/>
- Profinder. (n.d.). *Miten voimme auttaa?*. Noudettu 2020-5-1 osoitteesta <https://profinder.fi/palvelut/>
- Rajakannas, V. (2018a). *Platform Companies Won. What Comes Next for Finnish Companies?* OP Financial Group. Noudettu 2020-2-19 osoitteesta <https://op-developer.fi/articles/platform-companies-won-what-comes-next-for-finnish-companies>
- Rajakannas, V. (2018b). *The Winning Circle of Platform Companies is Based on Data*. OP Financial Group. Noudettu 2020-6-26 osoitteesta <https://op-developer.fi/articles/the-winning-circle-of-platform-companies-is-based-on-data>
- Runkler, T. A. (2016). *Data Analytics. Models and Algorithms for Intelligent Data Analysis* (2nd edition). Springer Vieweg.

- Ruokavirasto. (2019). *Avoim tieto*. Noudettu 2020-4-5 osoitteesta <https://www.ruokavirasto.fi/tietoa-meista/avointieto/>
- Russom, P. (2011). *Big Data Analytics*. TDWI Research. Noudettu 2020-3-5 osoitteesta <https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf>
- Schatsky, D., Camhi, J. & Muraskin, C. (2019). *Data Ecosystems. How Third-Party Information Can Enhance Data Analytics*. Deloitte. Noudettu 2020-3-21 osoitteesta <https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html#endnote-sup-3>
- Sun, Z. & Huo, Y. (2019). The Spectrum of Big Data Analytics. *Journal of Computer Information Systems*. <https://doi.org/10.1080/08874417.2019.1571456>
- Suomen Pankki. (n.d.). *Suomen Pankin ja Finanssivalvonnan avoimen datan palvelu*. Noudettu 2020-4-5 osoitteesta <https://www.suomenpankki.fi/fi/suomenpankki/yleisopalvelut/suomen-pankin-ja-finanssivalvonnan-avoimen-datan-palvelu/>
- SYKE. (2020a). *Avoim tieto*. Suomen ympäristökeskus. Noudettu 2020-4-5 osoitteesta <https://www.syke.fi/avointieto>
- SYKE. (2020b). *Suomen avoin meridata*. Suomen ympäristökeskus. Noudettu 2020-5-15 osoitteesta <https://itameri.fi/fi-FI/Datapalvelut>
- Tampere. (n.d.). *Tampereen kaupungin dataportaali. Avointa dataa Tampereen seudulta*. Tampereen kaupunki. Noudettu 2020-4-4 osoitteesta <https://data.tampere.fi/fi/>

- Telia. (2018). *Crowd Insights auttaa tekemään parempia päätöksiä. Ymmärrä väkijoukkojen liikkeitä ja tee dataan perustuvia ratkaisuja*. Telia Company AB. Noudettu 2020-4-19 osoitteesta <https://www.telia.fi/yrityksille/palvelut/teknologiat/crowd-insights>
- THL. (2020). *Avoin data*. Terveyden ja hyvinvoinnin laitos. Noudettu 2020-4-5 osoitteesta <https://thl.fi/fi/tilastot-ja-data/aineistot-ja-palvelut/avoin-data>
- Tilastokeskus. (n.d.). *Avoimet tietokanta-aineistot*. Noudettu 2020-4-4 osoitteesta <http://www.stat.fi/org/avoindata/pxweb.html>
- Traficom. (2020). *Avoin data*. Liikenne- ja viestintävirasto. Noudettu 2020-4-4 osoitteesta <https://www.traficom.fi/fi/tilastot-ja-julkaisut/avoin-data>
- Traffic Management Finland. (n.d.). *Digitraffic ja avoin data*. Noudettu 2020-5-6 osoitteesta <https://www.tmf.fi/fi/tmf/avoindata>
- Tukes. (n.d.). *Avoin data*. Turvallisuus- ja kemikaalivirasto. Noudettu 2020-4-5 osoitteesta <https://tukes.fi/tietoa-tukesista/avoin-data>
- Tukey, J. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1), 1-67. Noudettu 2020-6-22 osoitteesta [www.jstor.org/stable/2237638](http://www.jstor.org/stable/2237638)
- Tutkihankintoja.fi. (n.d.). *Tutki hankintoja*. Hansel Oy. Noudettu 2020-6-17 osoitteesta <https://tutkihankintoja.fi>
- Vainio, N. (2011). *Vapaus ja yhteisöllisyys vapaiden ohjelmistojen liikkeenlähtökohtina*. Tampere University Press. [http://urn.fi/URN:ISBN 978-951-44-8427-8](http://urn.fi/URN:ISBN%20978-951-44-8427-8)



Valtiovarainministeriö. (n.d.). *Avoin tieto*. Noudettu 2020-3-18 osoitteesta <https://vm.fi/avoin-tieto>

Varteva, A. & Suvanto, N. (2017). Avoin data kutsuu, tartu tilaisuuteen! *Talouselämä*. Noudettu 2020-2-25 osoitteesta <https://www.talouselama.fi/kumppaniblogit/6aika-avoin-data-ja-rajapinnat/avoin-data-kutsuu-tartu-tilaisuuteen/d942188f-ce36-3f15-a9ba-5f76486cbbf2>

Verhoeven, R., S. (2017). Capitalizing on External Data is Not Only an Outside-In Concept. *Compact*, 1(2017), 25-30. Noudettu 2020-3-21 osoitteesta <https://www.compact.nl/en/articles/capitalizing-on-external-data-is-not-only-an-outside-in-concept/#ref>

Väylä. (2019). *Avoin data*. Väylävirasto. Noudettu 2020-4-4 osoitteesta <https://vayla.fi/avoindata>

Ward, J. S. & Barker, A. (2013). *Undefined By Data: A Survey of Big Data Definitions*. University of St Andrews. Noudettu 2020-2-28 osoitteesta <https://arxiv.org/pdf/1309.5821.pdf>

Woods, D. (2012). *Do You Suffer from the Data Not Invented Here Syndrome?* Forbes. Noudettu 2020-3-22 <https://www.forbes.com/sites/danwoods/2012/11/15/do-you-suffer-from-the-data-not-invented-here-syndrome/#228415e44d5c>