TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

KRISTIINA HEINONEN

AVOIMEN DATAN HYÖDYNTÄMISEN HAASTEET GLOBAALIN ORGANISAATION BIG DATA ANALYTIIKASSA

Diplomityö

Tarkastaja: professori Samuli Pekkola
Tarkastaja ja aihe hyväksytty
Talouden ja rakentamisen tiedekunta-
neuvoston kokouksessa 9.12.2015

TIIVISTELMÄ

KRISTIINA HEINONEN: Avoimen datan hyödyntämisen haasteet globaalin organisaation big data analytiikassa
Tampereen teknillinen yliopisto
Diplomityö, 67 sivua, 9 liitesivua
Maaliskuu 2016
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
Pääaine: Tuote- ja prosessitiedon hallinta
Tarkastaja: professori Samuli Pekkola

Avainsanat: Avoin data, Big data, Data-analytiikka, Globaali organisaatio, Hyödyntäminen, Haasteet

Useat organisaatiot ovat viime vuosina ryhtyneet avaamaan dataansa kansalaisten, tutkijoiden ja muiden organisaatioiden sekä sidosryhmien vapaasti saataville. Datan avaaminen voi auttaa monimutkaisten ongelmien ratkaisussa sekä uusien innovaatioiden syntymisessä, mutta myös antaa tukea päätöksentekoon ja osallistuttaa suurempia määriä ihmisiä tietojen analysointiin sekä lisätä läpinäkyvyyttä julkisten organisaatioiden toiminnassa. Avoimelle datalle ei ole vielä olemassa vakiintunutta määritelmää, mutta usein avoimella datalla tarkoitetaan dataa, johon kuka tahansa voi avoimesti päästä käsiksi, käyttää, muokata ja jakaa sitä mihin tahansa käyttötarkoitukseen. Big data on avoimen datan tavoin hyvin uusi ilmiö, jolla tarkoitetaan erityisen suurten ja järjestämättömien tietomassojen keräämistä, säilyttämistä ja ennen kaikkea analysointia tietoteknisten ratkaisujen avulla. Big data analytiikka nähdään siten työkaluna erityisen suurten datamäärien käsittelyyn ja analysointiin.

Tässä tutkimuksessa keskityttiin hyvin suurten ja sisällöltään vaihtuvien avoimien data-aineistojen tarkasteluun, jolloin puhutaan avoimesta big datasta. Tutkimusongelmana oli tietämättömyys siitä, mitä haasteita esiintyy avoimen datan hyödyntämisessä osana globaalin organisaation big data analytiikkaa. Tutkimus toteutettiin tarkastelemalla loppukäyttäjän näkökulmasta 16 avoimen datalähteen saatavuutta, kokonaisuutta ja laatua, dataformaatin avoimuutta ja koneluettavuutta, teknistä avoimuutta arkkitehtuurien ja rajapintojen avulla, käyttöehtoja ja uudelleenkäytettävyyden mahdollisuutta, kustannuksia, sekä datan ymmärrettävyyttä metadatan kuvaamista. Tämän lisäksi tarkasteltiin data-aineiston maantieteellistä kattavuutta, havaintotarkkuutta sekä ajanjaksoa, jolta avoin data oli kerätty. Tutkimus toteutettiin laadullisena dokumenttisanalyysinä, joka hyödyntää eksploratiivisen tutkimuksen strategiaa.

Tutkimuksen keskeisinä havaintoina määriteltiin seitsemän haastetta: tiedostomuotojen eroavaisuudet, puutteet metadatassa, erot havaintotarkkuuksissa, maantieteelliset rajoitteet, heikko arkkitehtuurikuvaus ja rajapinnat, eroavaisuudet datan laadussa sekä heikko saatavuus ja löydettävyys. Avoimien datalähteiden yhdisteleminen on haastavaa ja työlästä eikä aineistojen sisältämää dataa kuvata usein tarpeeksi tarkalla tasolla. Aineistot eivät myöskään usein ole sellaisenaan yhteensopivia toistensa kanssa ja eri aineistojen yhdisteleminen tuottaa paljon manuaalista työtä. Ratkaisuvaihtoehtona avoimen datan yhtenäistämiselle tutkimus esittää erillisen avoimen datan standardin määrittämistä. Standardin tulisi pitää sisällään yksiselitteinen määritelmä avoimelle datalle sekä ehdot sille, miten avointa dataa tulisi tarjota ja avata uudelleenhyödynnettäväksi.

ABSTRACT

KRISTIINA HEINONEN: Challenges of Utilizing Open Data in Global Organizations Big Data Analytics
Tampere University of Technology
Master of Science Thesis, 67 pages, 9 Appendix pages
March 2016
Master's Degree Programme in Information and Knowledge Management
Major: Product and Process Information Management
Examiner: Professor Samuli Pekkola

Keywords: Open data, Big data, Data analytics, Global organization, Utilization, Challenges

In recent years, several different organizations have started to open their data available to citizens, researches, other organizations and stakeholders. Opening data can help organizations to solve complex problems as well as emerge new innovations, provide support for decision-making and engage larger amount of people to analyze data. Data opening can also increase the transparency of the public organizations activities. There is still no standard definition for open data, but often open data as a term refers to data that any person may freely access, use, modify and distribute for any purpose. Like open data, big data is also very new phenomenon. Big data refers to extremely large amount of structured or non-structured data; it's collection, storage and analysis utilizing information technology solutions. Big data analytics can be seen as advanced analytic techniques operate on big data sets.

This study focused on examination of extremely large and varying type of open datasets, more specifically open big data. The research problem of the study was the lack of awareness of what are the challenges of utilizing open data as a part of global organizations big data analytics. The study was conducted by exanimating the availability of 16 different open data sources and their availability, completeness and quality, data format, architectures and interfaces, terms of use and re-usability, costs and descriptions of metadata. Also the study examined data sources geographical coverage, observation accuracies and time periods. The study was conducted as qualitative document analysis and utilized strategy of explorative research.

As key findings, the study identified seven challenges of utilizing open data in global organizations big data analytics: the differences in data formats, the lack of metadata descriptions, differences in observation accuracies, geographical constraints, weak descriptions of architectures and interfaces and differences in the quality of data as well as poor access and findability. Combining different open data sets is challenging and laborious and particularly the data is not describes sufficiently detailed level. Different data sets are not often compatible as such and combining data sets usually produces a lot of manual work. Defining a separate open data standard could be a solution for harmonizing different open data sets. The standard should specify a clear definition of open data but also determine the terms and conditions how open data should be opened on provided for users to reuse.

ALKUSANAT

Ajatus tutkimuksen aiheesta sai alkunsa Tampereen teknillisen yliopiston tiedonhallinnan ja logistiikan laitoksen tutkimuskeskus NOVI:n hankkeesta, jossa tutkittiin ulkoisten datalähteiden saatavuutta ja hyödyntämismahdollisuuksia tuulivaihteita valmistavan globaalin yrityksen liiketoiminnassa. Hankkeen tutkimusongelmana oli selvittää, miten erilaisista ulkoisista tietolähteistä saatavaa dataa voidaan hyödyntää tuulivoimaloiden tuulivaihteiden huoltotoiminnan suunnittelussa ja varastojen optimoinnissa. Hankkeessa kartoitettiin avoimia datalähteitä, joita oli tarkoitus yhdistää globaalin organisaation sisäisiin datalähteisiin ja kartoittaa big data analytiikan avulla korrelaatioita sekä muita mahdollisia suhteita hyvin laajasta data-aineistosta. Tavoitteena oli big data analytiikan avulla selvittää, miksi tuulivaihteilla on hyvin erilaiset elinkaaret ja mitkä tekijät johtavat niiden vikaantumiseen. Tutkimusassistenttina hankkeessa toimiessani havaitsin useita avoimen datan hyödynnettävyyden haasteita ja motivaatio tämän tutkimuksen toteuttamiselle sai alkunsa. Diplomityössäni halusin keskittyä tunnistamaan avoimen datan hyödynnettävyyden haasteita erityisesti globaalin organisaation big data analytiikan näkökulmasta, jotta haasteet osattaisiin huomioida ja ennakoida tarkemmin tulevaisuudessa.

Erityisesti haluan kiittää professoriani Samuli Pekkolaa ajatuksista ja ohjeistuksesta diplomityöprosessin edetessä sekä mahdollisuudesta tehdä tutkimus hyvin mielenkiintoisesta ja ajankohtaisesta aiheesta Tampereen teknillisen yliopiston tiedonhallinnan ja logistiikan laitokselle. Haluan myös kiittää nykyistä työnantajaani Deloitte Oy:ta ja erityisesti kollegoitani saamastani tuesta ja opeista prosessin aikana, joiden myötä olen päässyt käytännössä myös hyödyntämään diplomityöni tuloksia ja opiskelun myötä saatuja oppeja liikkeenjohdon konsulttina. Viimeisenä haluan kiittää myös perhettäni sekä avopuolisoani tuesta yhteensä kahdeksan kuukautta kestäneen diplomityöprosessin aikana, jolloin päivän tunnit eivät tuntuneet riittävän kaiken tarvittavan työn tekemiseen. Tutkimuksen toteutus on tarjonnut niin pitkiä iltoja kuin hienoja elämyksiä ja oivalluksia.

Tampereella, 21.3.2016

Kristiina Heinonen

SISÄLLYSLUETTELO

KUVA- JA TAULUKKOLUETTELO

KESKEISET KÄSITTEET

Data	Digitaalisesti tallennettua informaatiota, kuten dokumentteja, tietokantoja ja tallenteita, eli raaka-ainetta, josta voidaan jalostaa merkityksellisempää informaatiota. (Poikola et al. 2010, s. 14)
Informaatio	Informaatio on datan ihmiselle tuottama mielle tai merkitys. (Borglund & Engvall 2014, s. 167)
Avoin data	Avoimella datalla tarkoitetaan dataa, johon kuka tahansa voi avoimesti päästä käsiksi, käyttää, muokata ja jakaa sitä mihin tahansa käyttötarkoitukseen. Datan tarjoajana voi toimia joko julkinen tai yksityinen organisaatio, ja se on saatavilla koneluettavassa muodossa ilman teknisiä rajoitteita ja erillisiä kustannuksia. (Dietrich et al. 2015)
Big data	Big datalla viitataan hyvin laajoihin data-aineistoihin, joiden koko on suurempi kuin tyypilliset tietokantaohjelmistot pystyvät tallentamaan, hallitsemaan ja analysoimaan tehokkaasti. (Chui et al. 2011, s. 1)
Edistyksellinen analytiikka	Edistyksellisellä analytiikalla tarkoitetaan teknologioiden hyödyntämistä vastaamaan datasta haettavien kysymysten tai ongelmien ratkaisuun. Edistyksellinen analytiikka ei itsessään ole teknologia, vaan joukko työkaluja, joita käytetään yhdessä datan analysointiin ja ennustamaan tuloksia ongelmien ratkaisuun. (Bose 2009, s.156)
Big data analytiikka	Big data analytiikalla tarkoitetaan edistyksellisten analytiikkatyökalujen käyttöä big datan jäsentämisessä. (Russom 2011, s. 8)
Organisaatio	Huolellisesti suunniteltu järjestelmä, jonka tehtävänä on toteuttaa sille asetetut tavoitteet (Harisalo 2008), kuten yritys, hallinto tai järjestö.
Globaali organisaatio	Globaalilla organisaatiolla viitataan kansainvälisesti toimivaan organisaatioon, joka toimii useissa eri maissa.

1. JOHDANTO

Viime vuosina useat tutkimukset ovat osoittaneet datan avaamisen tuovan etuja kansalaisille, tutkijoille, organisaatioille ja muille sidosryhmille, kun data-analytiikan avulla on voitu ymmärtää ongelmia uusin tavoin. (Zuiderwijk et al. 2014; Janssen et al. 2012) Yhdistelemällä data-aineistoja uudella tavalla voidaan luoda uutta tietoa sekä tehdä oivalluksia, jotka synnyttävät kokonaan uusia palveluita tai aluevaltauksia. (Dietrich et al. 2015) Yhä useammat organisaatiot ympäri maailman ovatkin alkaneet avaamaan dataansa vapaasti kaikkien hyödynnettäväksi (Zuiderwijk et al. 2012, s. 167) ja useiden menestystarinoiden myötä on nähtävissä, miten avoimen datan avulla kehitetyt sovellukset ja palvelut ovat synnyttäneet uusia innovaatioita ja tuottaneet uusia liiketoimintamahdollisuuksia.

Avoimen datan on ennustettu olevan uusi kultakaivos, joka vauhdittaa innovaatioita ja tuottaa valtavia tuloja. (The Economist 2013) McKinseyn tekemän tutkimuksen mukaan avoin data mahdollistaa globaalisti jopa yli 2,5 biljoonan euron lisäarvon esimerkiksi tuottavuuden kasvun, uusien tuotteiden ja palveluiden sekä toiminnan tehostamisen myötä. (Manyika et al. 2013, s. 2) Myös Euroopan komissio on selvityksessään todennut, että Euroopan unioni voi saada vuosittain lähes 40 miljardia euroa taloudellisia hyötyjä avaamalla dataa uusiokäyttöä varten. Datan avaaminen tukee myös politiikan eri osa-alueita, kuten ympäristöpolitiikkaa ja edistää kansalaisten osallistumista poliittiseen sekä yhteiskunnalliseen toimintaan. (Euroopan komissio 2011) Tällä hetkellä suurimman osan avoimesta datasta tarjoavat julkiset organisaatiot, jotka haluavat toiminnalleen läpinäkyvyyttä (Janssen 2011, s. 446) ja toteuttavat avoimen hallinnon periaatteita (Ubaldi 2013, s. 4).

Saatavilla oleva avoin data vaihtelee sää- ja paikkatiedoista julkisen sektorin budjetteihin ja liikennetietoihin (Janssen et al. 2012, s. 258) ja usein suurin hyöty saadaan, kun avointa dataa yhdistetään organisaation sisäisiin data-aineistoihin. (Manyika et al. 2013, s. 7) Kun puhutaan hyvin suurista ja jäsentymättömistä datamääristä, liitetään keskusteluun usein big datan ilmiö. Peltolan (2014) mukaan avoin data on yksi parhaista big datan raaka-aineista. Erityisesti big data analytiikan avulla on mahdollista hyödyntää suuria määriä reaaliaikaista ja historiallista tietoa sekä löytää uusia malleja, poikkeavaisuuksia tai korrelaatioita eri datalähteiden väliltä. Avoin data syventää entisestään big data analytiikan mahdollisuuksia (Manyika et al. 2013, s. 1), kun ulkoisten datalähteiden saatavuus lisääntyy.

Hellbergin & Hedströmin (2014, s. 46) mukaan avoimesta datasta puhuttaessa näyttää olevan hyvin tavallinen olettamus, että jos dataa avataan kaikkien käytettäväksi, ihmiset

alkavat automaattisesti hyödyntää sitä. Suurimpana haasteena onkin, että avoimella datalla ei sellaisenaan ole arvoa; siitä tulee arvokasta vasta kun sitä käytetään. (Janssen et al. 2012) Käyttäjät tulisi saada kiinnostumaan saatavilla olevasta datasta ja sen hyödyntämismahdollisuuksista tekemällä siitä mahdollisimman helppoa ja vaivatonta siten, että kynnys avoimen datan hyödyntämiselle olisi mahdollisimman matala.

Tutkimukset kuitenkin osoittavat, että datan avaamisesta huolimatta on sen tehokkaalle hyödyntämiselle löydettävissä myös haasteita ja esteitä (Zuiderwijk et al. 2012; Janssen et al. 2012; Huijboom & Broek 2011). Avoimen datan hyödyntäminen saattaa usein vaatia erillisten ohjelmistojen käyttöä sekä teknistä ymmärrystä aiheesta. Dataa on myös satavilla useissa eri formaateissa ja muodoissa, jonka vuoksi se kärsii standardisoinnin puutteesta. (Janssen et al. 2012) Tämä tutkimus on toteutettu empiirisesti kartoittamalla saatavilla olevia avoimia datalähteitä ja näiden hyödynnettävyyden haasteita erillisen mittariston avulla. Tutkimustuloksena työ esittelee löydettyjä haasteita globaalin organisaation big data analytiikan näkökulmasta ja pyrkii pohtimaan ratkaisuvaihtoehtoja haasteiden välttämiseksi tulevaisuudessa.

1.1 Tutkimuksen lähtökohdat ja tausta

Tutkimuksen lähtökohtana toimi Tampereen teknillisen yliopiston tiedonhallinnan ja logistiikan laitoksen tutkimuskeskus NOVI:n hanke, jossa tutkittiin ulkoisten datalähteiden saatavuutta ja hyödyntämismahdollisuuksia tuulivaihteita valmistavan globaalin yrityksen liiketoiminnassa. Hankkeen tutkimusongelmana oli selvittää miten erilaisista ulkoisista tietolähteistä saatavaa dataa voidaan hyödyntää tuulivoimaloiden tuulivaihteiden huoltotoiminnan suunnittelussa ja varastojen optimoinnissa. Tavoitteena oli myös kehittää malleja ja suosituksia siihen miten tällainen data on kerättävissä ja hyödynnettävissä tulevaisuudessa yrityksen big data analytiikassa ja edelleen liiketoiminnan kehittämisessä. Ulkoisten datalähteiden kartoituksessa keskityttiin suurimmaksi osaksi avoimien datalähteiden tarkasteluun.

Hankkeessa hyödynnettiin ulkoisten datalähteiden lisäksi yrityksen sisäisiä tietolähteitä ja etsittiin näiden väliltä korrelaatioita sekä muita mahdollisia suhteita erilaisia louhintamenetelmiä hyväksikäyttäen. Tavoitteena oli big data analytiikan avulla selvittää, miksi tuulivaihteilla on hyvin erilaiset elinkaaret ja mitkä tekijät johtavat niiden vikaantumiseen. Ulkoisina datalähteinä pyrittiin käyttämään apuna esimerkiksi avoimesti saatavilla olevaa sää- ja paikkatietoa sekä tuuliturbiineista ja sähköverkoista saatavaa tietoa. Tavoitteena oli, että ulkoiset datalähteet kattavat maailmanlaajuisen tarkastelun vähintään viimeisen 15 vuoden ajalta.

Hankkeessa havahduttiin jo alkuvaiheessa avoimen datan hyödyntämisen haasteisiin, kun tavoitteena oli yhdistää avointa dataa useista ulkoisista tietolähteistä globaalisti toimivan organisaation tarpeisiin. Ulkoisten datalähteiden kartoituksessa löydettiin useita hyödyllisiä avoimen datan lähteitä, mutta niiden yhdistäminen ja tehokas hyödyntä-

minen eivät osoittautuneet kovinkaan yksinkertaiseksi tehtäväksi. Hankkeessa havaittujen ilmiöiden pohjalta avoimen datan hyödyntämisen haasteisiin globaalin organisaation big data analytiikassa keskityttiin tarkemmin tässä tutkimuksessa, jotta haasteet osattaisiin huomioida ja ennakoida tarkemmin tulevaisuudessa.

1.2 Tutkimusongelma ja tutkimuksen tavoitteet

Tutkimusongelman määrittelyn lähtökohtana toimivat tutkimuksen taustan asettamat vaatimukset tutkimuksen toteutukselle. Tutkimusongelmaksi voidaan jäsentää tietämättömyys siitä, mitä haasteita esiintyy avoimen datan hyödyntämisessä osana globaalin organisaation big data analytiikkaa. Tutkimusongelma jakautuu päätutkimuskysymykseen ja sitä tukeviin alatutkimuskysymyksiin. Alatutkimuskysymykset eivät siis sisällä uusia näkökulmia, vaan ainoastaan selventävät ja täsmentävät päätutkimusongelmaa. Päätutkimuskysymys voidaan määritellä seuraavasti:

- Mitä haasteita esiintyy avoimen datan hyödyntämisessä globaalin organisaation big data analytiikassa?

Päätutkimuskysymykseen pyritään vastaamaan seuraavien alatutkimuskysymysten avulla. Alatutkimuskysymyksiä ovat:

- Miten avoimen datan hyödynnettävyyttä voidaan mitata?
- Millaisia haasteita tai esteitä avoimen datan käytössä on organisaatiolle?
- Miten avointa dataa voidaan hyödyntää organisaation big data analytiikassa?

Alatutkimuskysymysten avulla pyritään täsmentämään, miten avoimen datan hyödynnettävyyttä voidaan ylipäätään mitata sekä millaisia haasteita tai esteitä avoimen datan käytölle on tunnistettu organisaatioissa. Alatutkimuskysymysten avulla pyritään myös kuvaamaan tutkimuksen kahden keskeisen datatyypin, avoimen datan ja big datan, välistä suhdetta. Ensimmäinen alatutkimuskysymys keskittyy määrittelemään menetelmiä ja kriteerejä, joiden avulla aineiston hyödynnettävyyden mittaaminen on mahdollista. Toinen alatutkimuskysymys kartoittaa millaisia haasteita tai esteitä avoimen datan käytölle on organisaatioissa jo tunnistettu. Viimeinen alatutkimuskysymys keskittyy tarkastelemaan avoimen datan ja big datan suhdetta sekä kartoittamaan avoimen datan hyödyntämisenmahdollisuuksia osana organisaation big data analytiikkaa. Alatutkimuskysymyksiin pyritään vastaamaan pääosin tutkimuksen teoriakatsauksen avulla.

Tutkimuksen tavoitteena on löytää vastaus esitettyyn tutkimusongelmaan edellä mainittujen tutkimuskysymysten avulla ja siten kartoittaa avoimen datan hyödyntämiseen liittyviä haasteita globaalin organisaation big data analytiikassa, jotta haasteet tiedostettaisiin nykyistä paremmin. Perimmäisenä tavoitteena on, että löydettyihin haasteisiin pystyttäisiin löytämään myös ratkaisuja tulevaisuudessa. Tutkimuksen tavoitteena on myös luoda katsaus sekä olemassa olevaan teoriaan avoimen datan hyödyntämisestä ja big

data analytiikasta, että kartoittaa millaisia haasteita avoimen datan hyödyntämisessä on jo havaittu. Työn tarkoituksena on verrata näitä jo löydettyjä haasteita empiirisen tutkimuksen avulla löydettyihin haasteisiin, jotka esiintyvät erityisesti globaalin organisaation big data analytiikan näkökulmasta.

1.3 Tutkimuksen rajaus

Tutkimus on rajattu tarkastelemaan avoimen datan hyödyntämisen haasteita erityisesti globaalisti toimivan organisaation näkökulmasta, joka hyödyntää avointa dataa osana big data analytiikkaa. Globaalisti toimiva organisaatio eroaa rajoittuneemmalla maantieteellisellä alueella toimivista organisaatioista siten, että myös avointa dataa halutaan hyödyntää mahdollisimman laajalta tarkastelualueelta. Saatavilla oleva avoin data on lähes poikkeuksetta maantieteellisesti rajoitettua, minkä vuoksi globaalit organisaatiot joutuvat usein yhdistelemään data-aineistoja monista eri lähteistä.

Avointa dataa ja sen mahdollistamia hyötyjä voidaan tarkastella sekä datan tuottajan että loppukäyttäjän näkökulmasta. Tässä tutkimuksessa avoimen datan hyödyntämisen haasteita tarkastellaan loppukäyttäjän näkökulmasta, jonka tarkoituksena on hyödyntää avointa dataa organisaation sisäisen datan tukena. Tutkimusaineistona olevat avoimet datalähteet rajoittuvat myös tarkastelemaan lähinnä sää- ja paikkatietokantoja sekä tuuliturbiineista ja sähköverkoista saatavia avoimen datan lähteitä.

Tutkimus toteutetaan laadullisena dokumenttisanalyysinä, joka hyödyntää eksploratiivisen tutkimuksen strategiaa. Tutkimuksen taustalla on siten nähtävissä hermeneuttinen tieteenkäsitys, sillä tutkimuksessa korostuvat tutkijan oma tulkinta, käsitys sekä ymmärrys aiheesta. Tutkimuksen metodologia sekä tutkimusmenetelmät on kuvattu tarkemmin luvussa 4. Tutkimuksen aihepiiri on hyvin uusi, joka osaltaan aiheuttaa myös rajoitteita tutkimuksen saatavilla olevalle aineistolle. Aineisto on rajattu vain tiettyyn otokseen saatavilla olevista avoimista datalähteistä.

1.4 Tutkimuksen rakenne

Tutkimuksen rakenne jakautuu neljää osaan: esittelyyn, teoria- ja empiriaosuuksiin sekä päätelmiin. Ensimmäisen osan tavoitteena on esitellä tutkimuksen lähtökohdat, kun taas toinen osa keskittyy kuvaamaan tutkimuksen teoriataustaa niin avoimen datan kuin big data analytiikankin näkökulmista. Kolmas osa sisältää itse empiirisen tutkimuksen toteutuksen ja tulosten esittelyn. Viimeinen osa puolestaan esittelee päätelmät tutkimustulosten pohjalta. Tutkimuksen rakenne on esitelty tarkemmin kuvassa 1.

Kuva 1. *Tutkimuksen rakenne*

Työn ensimmäisessä luvussa kuvataan tutkimuksen taustaa ja lähtökohtia sekä määritellään tutkimusongelma. Luvussa esitellään myös motivaatio tutkimuksen taustalla. Tutkimusongelma kuvataan asetettujen tutkimuskysymysten avulla, sekä määrittämällä rajaukset ja linjaukset tutkimuksen toteuttamiselle.

Työn teoriaosuus koostuu luvuista kaksi ja kolme. Toinen luku keskittyy avoimeen dataan, sen käsitteen määrittelyyn sekä suhteen muihin datatyyppeihin. Luvussa esitellään myös erilaisia avoimen datan hyödynnettävyyden mittareita. Luku sisältää myös avoimen datan prosessimallin kuvauksen sekä esittelee hyötyjä ja esteitä avoimen datan käytölle. Kolmas luku puolestaan keskittyy big data analytiikkaan. Luku esittelee big datan määritelmän, sekä pohtii miten edistyksellistä analytiikkaa voidaan hyödyntää big datan käsittelyyn. Luku jäsentää myös, miten big data analytiikkaa voidaan hyödyntää liiketoiminnassa ja millaisia työkaluja analytiikan tukena voidaan käyttää. Tämä lisäksi luku esittelee big datan ja avoimen datan suhdetta sekä määrittelee avoimen big data analytiikan käsitteen. Viimeisenä luku esittelee koosteen tutkimuksen teoriasta ja ilmentää, miten avoimen datan ja big data analytiikan teoriataustaa on hyödynnetty työn empiirisessä osuudessa.

Tutkimusmenetelmät ja aineisto kuvataan neljännessä luvussa. Tehdyt tutkimusmetodologiset valinnat sekä tieteenkäsitys ja tutkimusote kuvataan luvussa tarkemmin. Myös valittu tutkimusmenetelmä ja aineistonkeruu sekä analysointi esitellään. Viidennessä luvussa kuvataan työn aineiston pohjalta saadut tutkimuksen varsinaiset tulokset, eli

haasteet, jotka avoimen datan hyödyntämisessä osana globaalin organisaation big data analytiikkaa tulee ottaa huomioon.

Kuudennessa luvussa arvioidaan saatuja tutkimustuloksia, sekä peilaa niitä työn teoriaosuuteen. Luvussa esitellään myös ratkaisuvaihtoehtoja löydettyjen haasteiden välttämiseksi ja lopuksi pohditaan tutkimuksen puutteita, onnistumisia sekä jatkotutkimuksen tarvetta. Työn viimeinen luku sisältää yhteenvedon koko työn sisällöstä ja toteutetun tutkimuksen kulusta.

2. AVOIN DATA

Avoimen datan juuret ovat Lindmanin et al. (2013, s. 1241) mukaan lähtöisin avoimen lähdekoodin trendistä (Fitzgerald 2006) sekä seurausta osaltaan myös avoimen innovoinnin ilmiöstä (Lichtenthaler 2011) ja tieteellisen tiedon ja tutkimustulosten niin sanotusta Open access -julkaisemisesta (Willinsky 2006). Open access -julkaisemisella tarkoitetaan, että tutkimustulokset ja käytetyt tausta-aineistot ovat tiedeyhteisön ja kaikkien muidenkin vapaasti saatavilla. Myös avoimen hallinnon periaate ja sen yleistyminen on osaltaan vaikuttanut suuresti keskusteluun datan avaamisen tärkeydestä (Ubaldi 2013; Lee & Kwak 2012; Dawes & Helbig 2010).

Terminä avoimuus on otettu käyttöön vasta 2000-luvun loppupuolella yleistämään käsitettä julkisesti ja helposti kaikkien saatavilla olevasta tiedosta. Avoimuus on yleistynyt myös koskemaan dataa, joka viittaa hyödyllistä tietoa sisältävään raakadataan, jota voidaan edelleen jäsentää, analysoida ja esittää eri muodoissa, joka puolestaan johtaa uuden tiedon syntymiseen. (Hoxha & Brahaj 2011) Vaikka avoimen data ilmiö on vasta Manyikan et al. (2013, s. 2) mukaan aluillaan, näkevät he sen tuovan merkittävää lisäarvoa, kun edistyksellisen analytiikan avulla avoimia ja organisaation sisäisiä datalähteitä voidaan yhdistellä. Avoimena datana voi olla saatavilla melkein mitä tahansa tietoa, joka on sallittua julkaista. Useimmiten se liittyy säähän, maantieteeseen, liikenteeseen, politiikkaan, julkisten organisaatioiden toimintaan ja budjetteihin sekä tilastoihin esimerkiksi ruuasta, sähkönkulutuksesta, koulutuksen tasosta tai turvallisuudesta (Janssen et al. 2012). Useimmiten dataa avataan tiettyä tarkoitusta varten, olipa tavoitteena vauhdittaa tutkimusta tai kehitystä, synnyttää uusia yrityksiä tai parantaa kansanterveyttä ja turvallisuutta (Gurin 2014a).

Saatavilla olevan avoimen datan lisääntymistä on kasvattanut merkitsevästi asetelma, jossa julkisille organisaatioille asetetaan painetta datan avaamiseksi. (Janssen et al. 2012; Zuiderwijk et al. 2012) Painostus datalähteiden avaamiselle voidaan jäljittää alun perin WWW:n kehittäjä Tim Berners-Leehen, joka julkaisi merkittävän tutkimuksen (Berners-Lee 2006) semanttisesta webistä ja avoimesta datasta. Tällä tutkimuksella on ollut myöhemmin huomattava vaikutus myös poliittisessa keskustelussa avoimen datan lisääntymisestä. (Lindman et al. 2013) Esimerkiksi Presidentti Obama antoi heti ensimmäisenä virkaanastumispäivänään kehotuksen lisätä hallinnon toiminnan avoimuutta ja myöhemmin samana vuonna asetti avoimen hallinnon direktiivin, jossa korostetaan kolmea periaatetta: avoimuus, osallistuminen ja yhteistyö. (Yhdysvaltojen presidentin kanslia 2009) Myös Euroopan komissio asetti vuonna 2013 direktiivin, jolla pyritään helpottamaan julkisen sektorin tuottaman avoimen datan käyttöä. (Euroopan komissio 2013) Samana vuonna G8 (2013) -maat allekirjoittivat avoimen datan peruskirjan, joka

määrittelee viisi periaatetta avoimelle datalle; vapaa ja avoin pääsy dataan on erittäin arvokasta yhteiskunnalle ja sen taloudelle, jonka vuoksi julkaistavan avoimen datan tulee olla laadukasta ja ajantasaista, se tulee olla yhtäläisesti kaikkien käytettävissä ja sen tärkeimpänä tavoitteena on hallinnon ja innovaatiotoiminnan parantaminen.

2.1 Avoin data käsitteenä

Kuten useiden uusien tietoteknologisten innovaatioiden ja tekniikoiden kanssa, myös avoimen datan käsitteen määrittelyssä ja sisällössä on eroavaisuuksia. (Tammisto & Lindman 2012, s. 297) Avoimen datan käsitettä on tarkasteltu tarkemmin useammissakin tutkimuksissa viime vuosina (Lindman et al. 2013; Borglund & Engvall 2014; Tammisto & Lindman 2012), mutta yhtä yksiselitteistä määritelmää ei ole onnistuttu tuottamaan. Avoimen datan käsitteen voidaan Borglundin ja Engvallin (2004, s. 163) mukaan ajatella olevan peräisin Euroopan unionin paikkatietoinfrastruktuurin (INSPIRE) perustamisesta vuodelta 2004. Sen tavoitteena oli tuottaa yhdenmukaista, laadukasta paikkatietoa yhteisön päätöksenteon täytäntöönpanoa, seurantaa ja arviointia varten. (Euroopan komissio 2004; Blakemore & Craglia 2006) Kun yhä enemmän tietoa tuotetaan digitaalisesti, mahdollisuus käyttää tietoa uudelleen kasvaa ja tästä usein puhutaan avoimena datana. (Borglund & Engvall 2014, s. 163)

Ubaldin (2013, s. 6) määritelmän mukaan avoin data voidaan käsittää datana, jota kuka tahansa voi käyttää ja hyödyntää vapaasti, käyttää uudelleen sekä jakaa siten, että enimmäisvaatimuksena on jakaa myös oma työnsä avoimesti muiden hyödynnettäväksi. Määritelmä on samankaltainen kuin yleisesti käytetty Open Knowledge Foundation -järjestön määritelmä, jonka mukaan data on avointa, kun kenellä tahansa on siihen vapaa pääsy, lupa käyttää, muokata ja jakaa sitä mihin tahansa tarkoitukseen. Avoimena datana tarjottavan aineiston tulee myös olla kokonaisuudessaan saatavilla enintään kohtuullisella kustannuksella, mieluiten maksutta ja ladattavissa internetin kautta. Aineiston pitää olla saatavilla käytännöllisessä ja muokattavassa muodossa, eli sen pitää olla vapaa teknisistä rajoitteista. (Open Knowledge Foundation 2015) Dietrich et al. (2015) täydentävät Open Knowledge Foundation -järjestön määritelmää korostamalla vielä avoimen datan tärkeintä ominaisuutta, eli yhteentoimivuutta muiden avoimien datalähteiden kanssa. Yhteentoimivuus on välttämätöntä, sillä sen avulla eri käyttäjät ja organisaatiot voivat toimia yhdessä ja rakentaa uusia, monimutkaisia järjestelmiä useiden avoimien datalähteiden pohjalta.

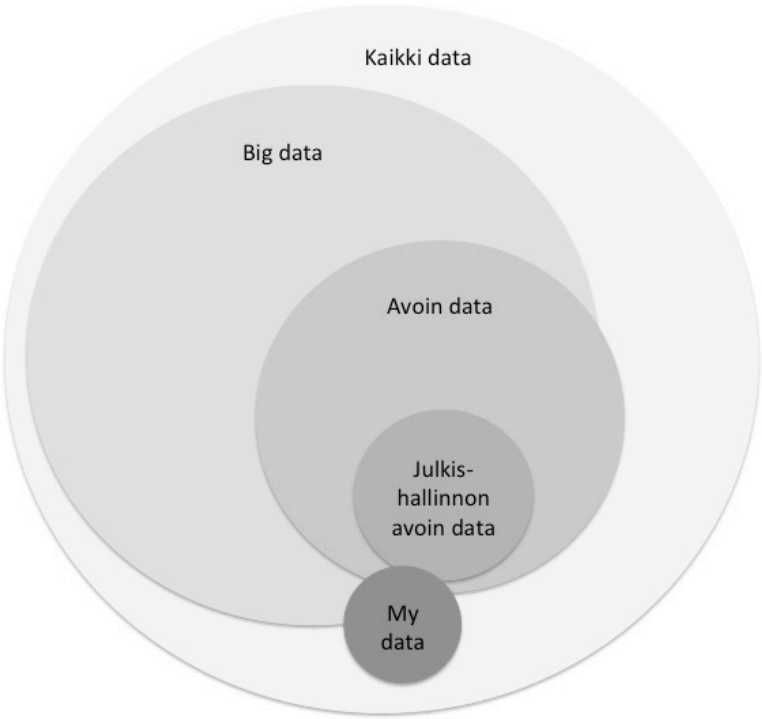
Useimmiten avoin data liitetään pelkästään julkisten organisaatioiden tuottamaksi (Halonen 2012, s. 12), sillä ne tuottavat tällä hetkellä selkeästi suurimman osan saatavilla olevasta avoimesta datasta (Janssen 2011). Tämän vuoksi osa määritelmistä pitää avointa dataa vain julkisten organisaatioiden tuottamana, kuten Janssen et al. (2012), joiden mukaan avoin data on julkista dataa, joka ei sisällä luottamuksellista tietoa ja on tuotettu julkisin varoin ilman minkäänlaisia käyttörajoitteita. Datan voi tarjota joko julkinen tai yksityinen yritys, kunhan data on rahoitettu julkisin menoin. Usein avoimen datan sy-

nonyymeinä käytetään myös käsitteitä julkisen sektorin data (eng. Public Sector Information) sekä julkishallinnon avoin data (eng. Open Government Data), jotka viittaavat molemmat vain julkishallinnon tuottamaan dataan. Borglund ja Engvall (2014, s. 164) huomauttavatkin, että avoimen datan termi ei ole vielä täysin vakiintunut, sillä siitä keskustellaan niin innovaatioiden, avoimen hallinnon ja julkisen talouden yhteydessä ilman, että tarkemmin määritellään mitä avoin data on.

Gurinin (2014) mukaan kaikkia avoimen datan määritelmiä yhdistää kuitenkin kaksi perusominaisuutta: tietojen on oltava julkisesti saatavilla ja kenen tahansa käytettävissä, sekä se on lisensoitu tavalla, joka mahdollistaa data uudelleenkäytön. Näin näkevät myös Borglund ja Engvall (2014, s. 164), joiden mukaan avoimen datan käsitteessä keskeisintä on, että data tuodaan julkiseksi ja tasapuolisesti kaikkien käytettäväksi.

2.2 Suhde muihin datatyyppeihin

Selkeimpiä esimerkkejä avoimen datan tuottajista ovat julkishallinnon organisaatiot. Manyikan et al. (2013, s. 4) mukaan avointa dataa ja julkishallinnon avointa dataa ei voida pitää synonyymeinä keskenään, vaan avoin data tulisi nähdä suurempana kokonaisuutena. Avoimet data-aineistot voivat olla lähtöisin julkishallinnolta, muilta toimielimiltä, yrityksiltä tai yksityishenkilöiltä, olivatpa ne suuria tai pieniä. Avoimen datan suhdetta muihin datatyyppeihin on kuvattu tarkemmin kuvassa 2.



Kuva 2. *Avoimen datan suhde muihin datatyyppeihin (mukaillen Manyika et al. 2013)*

Kaikella datalla viitataan yleisesti dataan, joka voidaan Poikolan et al. (2010, s. 14) mukaan määritellä informaatiotutkimuksen piirissä merkeistä ja symboleista koostuvana potentiaalisena informaationa. Data on digitaalisesti tallennettua informaatiota, kuten dokumentteja, tietokantoja ja tallenteita, eli raaka-ainetta, josta voidaan jalostaa merkityksellisempää informaatiota. Big data puolestaan viittaa hyvin laajoihin data-aineistoihin, joiden koko on suurempi kuin tyypilliset tietokantaohjelmistot pystyvät tallentamaan, hallitsemaan ja analysoimaan tehokkaasti (Chui et al. 2011, s. 1). Avoin data on useimmiten big dataa, mutta vain pieniä osuuksia big datasta voidaan tarjota avoimena datana. (Manyika et al. 2013, s. 4)

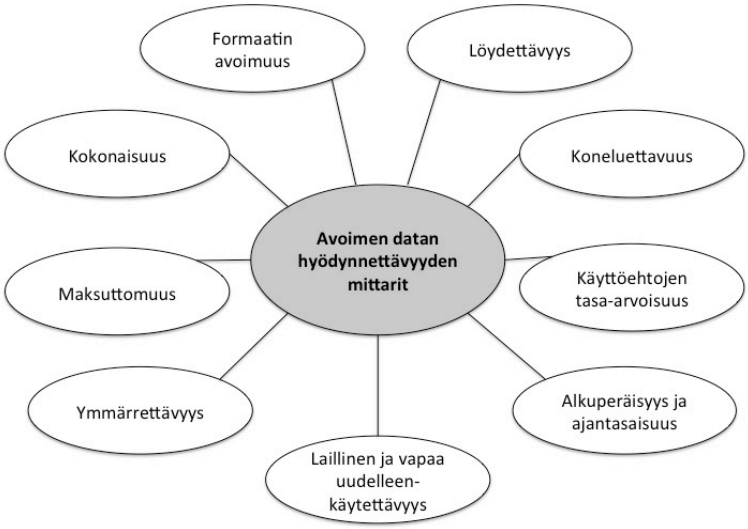
My datalla, eli suomennettuna omadatalla, viitataan Kuitusen et al. (2014, s. 4) mukaan ihmiskeskeiseen lähestymistapaan henkilötiedon hallinnassa ja käsittelyssä. Siinä ihmisille annetaan oikeus ja pääsy heistä kerättyyn dataan kuten ostotietoihin, liikennetietoihin, teletietoihin, terveystietoihin, taloustietoihin ja eri verkkopalveluihin kertyvään dataan. Keskeistä My datassa on, että ihmisille luodaan mahdollisuus siirtää tietojaan nykyistä uudelleenkäytettävämmässä muodossa itselleen tai valtuuttamaansa palveluun hyödynnettäväksi, esimerkiksi palveluiden räätälöintiä tai kohdistettua markkinointia varten. Loppuen lopuksi avoin data siis liittyy vahvasti My data -ilmiöön, jonka perustana on jakaa kerättyä dataa yksittäisten henkilöiden tai organisaatioiden hyödynnettäväksi (Manyika et al. 2013, s. 4).

2.3 Hyödynnettävyyden mittareita

Poikolan et al. (2010) mukaan avoimuuskeskustelussa puhutaan samaan aikaan muun muassa avoimesta lisensoinnista, teknisistä rajapinnoista, dataformaateista, metadatasta, tiedon harmonisoinnista, uudelleenkäytettävyydestä ja koneluettavuudesta. Täydellisesti hyödynnettävä avoin data sisältää siten hyvin monia eri ulottuvuuksia. Kaikki avoimuuden ulottuvuudet tulisi pyrkiä huomioimaan, jotta käyttäjällä olisi mahdollisuus hyödyntää avointa data-aineistoa mahdollisimman tehokkaasti. Käytännössä tilanne ei kuitenkaan vielä kovinkaan hyvin toteudu. (Susha et al. 2015, s. 182) Esimerkiksi teknisesti helposti saatavilla olevaan dataan saattaa usein kohdistua uudelleenkäyttöä tai uudelleenjakelua rajoittavia käyttöehtoja. Vastaavasti voi olla tapauksia, joissa muutoin täysin avoin data ei ole erityisen helposti hyödynnettävissä esimerkiksi hankalan tiedostomuodon, huonon löydettävyyden tai vajavaisen dokumentaation takia.

Ollakseen aidosti avointa, datan tulee olla avointa sekä teknologisesti, ideologisesti että laillisesti. (Halonen 2012) Kokonaisuudessaan avoimien data-aineistojen hyödynnettävyyttä voidaan Poikolan et al. (2010) mukaan arvioida seuraavien ulottuvuuksien mukaisesti: löydettävyys, kokonaisuus, käyttöehtojen tasa-arvoisuus, alkuperäisyys ja ajantasaisuus, laillinen ja vapaa uudelleen käytettävyys, maksuttomuus, koneluettavuus, formaatin avoimuus sekä ymmärrettävyys. Toinen jaottelu voidaan esittää Ubaldin (2013, s. 8) mukaan, joka määrittelee avoimen datan periaatteet seuraavasti: kokonaisuus, ensisijaisuus, ajantasaisuus, saatavuus, koneluettavuus, syrjimättömyys sekä pa-

tentoimattomuus ja lisenssivapaa, eli lähes identtisesti Poikolan et al. (2010) kanssa. Puolestaan Lindman et al. (2013) esittelevät avoimen datan ulottuvuudet tiiviimmin pelkästään teknisenä, oikeudellisena sekä kaupallisena avoimuutena. Onnistuminen datan avaamisessa vaatii siten paljon enemmän kuin pelkästään dataan käsiksi pääsyn mahdollistamisen. (Janssen et al. 2012) Jaotteluiden eroista huolimatta kaikki ulottuvuudet viittaavat sisällöiltään hyvin samoihin teemoihin ja hyödynnettävyyden kriteereihin, jotka on seuraavaksi esitelty tarkemmin Poikolan et al. (2010) tekemän jaottelun mukaisesti sekä esitetty havainnollisesti kuvassa 3.



Kuva 3. *Avoimen datan hyödynnettävyyden mittarit (mukaillen Poikola et al. 2010)*

2.3.1 Löydettävyys

Avoimen data-aineiston tulee olla helposti löydettävissä, eli sen olemassaolo ja sijainti tulee olla yleisesti tunnettu. Datan tulisi olla helposti löydettävissä internetistä sekä ihmisille että hakukoneille (Poikola et al. 2010), jotta käyttäjien on ylipäätään mahdollista päästä käsiksi tarjolla olevaan avoimeen dataan. Datan pitäisi myös olla löydettävissä ja saatavilla siten, että se voidaan ottaa käyttöön välittömästi ilman viranomaisen avustusta haluttuna ajankohtana (Open Knowledge Foundation 2015). Tällöin käyttäjän halutessa aineistoa voidaan hyödyntää esimerkiksi yöllä kello kolmelta.

Aineiston löydettävyyttä voidaan parantaa lisäämällä se esimerkiksi ylläpidettyyn datakatalogiin sekä optimoimalla datavarantojen kuvailutiedot hakukoneita varten sopiviksi (Poikola et al. 2010, s. 35). Useilla valtiolla on jo olemassa erillisiä dataportaaleita (Davies 2013), joiden avulla pyritään parantamaan avoimen datan saavutettavuutta. Suomessa alustana voidaan käyttää esimerkiksi Avoidata.fi -palvelua, joka vastaa Yhdysvaltojen Data.gov tai Britannian Data.gov.uk -palveluita. Kaikkien dataportaaleiden tarkoituksena on kerätä mahdollisimman paljon data-aineistoja samaan paikkaan. (Shadbolt & O'Hara 2013, s. 72; Lindman et al. 2013)

2.3.2 Kokonaisuus

Datan tulee olla kokonaisuudessaan vapaasti ladattavissa internetistä (Dietrich et al. 2015) eikä sen saavutettavuutta tai käyttömahdollisuuksia tule rajoittaa epäsuorasti, tarjoamalla esimerkiksi pääsy vain osaan tietokannasta kerrallaan. Usein avoimen datan kokonaisuutta rajoitetaan esimerkiksi erillisen kyselyrajapinnan kautta, jolloin aineistoa ei ole mahdollisuutta ladata kokonaisuudessaan. Kokonaisuuden rajoittamisella voidaan pyrkiä varmistamaan datan loukkaamattomuus ja estää rinnakkaisten kopioiden syntyminen. (Poikola et al. 2010, s. 35)

Kokonaisuudella viitataan myös osaltaan siihen, että kaikki data joka on mahdollista julkaista, tulisi tarjota avoimesti. (Halonen et al. 2012, s. 13) Esimerkiksi Ubaldi (2013, s. 8) korostaa kaiken julkisen datan avaamisen tärkeyttä, silloin kun kyseessä ei ole data joka voi vahingoittaa yksityisyyttä, turvallisuutta tai muita oikeuksia. Tällöin esimerkiksi verovaroin tuotetun datan tulisi olla kokonaisuudessaan kaikkien avoimesti hyödynnettävissä, jotta käyttäjien on mahdollista toteuttaa omat analyysinsä kerätystä datasta.

2.3.3 Käyttöehtojen tasa-arvoisuus

Avoimesti saatavilla olevan data-aineiston tulisi olla vapaa sosiaalisista ja organisatorisista rajoitteista siten, ettei henkilön työ, sijainti, asuinpaikka, organisaatiomalli, uskonto, poliittinen suuntautuneisuus tai etnisyys rajoita pääsyä dataan. (Open Knowledge Foundation 2015) Avoimen datan tulee olla tasa-arvoisesti kaikkien käytettävissä, oli käyttötarkoitus mikä tahansa. Halosen (2012, s. 13) mukaan datan käyttöoikeuksien rajoittaminen on moraalisesti väärin, sillä kaikilla tulisi olla yhtäläinen oikeus hyödyntää esimerkiksi verovaroin tuotettua dataa. Poikolan et al. (2010, s. 35) mukaan käytännössä tasa-arvoisuus toteutuu, mikäli aineisto on saatavissa verkosta ilman erillisiä rekisteröintivaatimuksia, jolloin kuka tahansa voi käyttää ja hyötyä datasta yhdenvertaisesti.

Avoimen datan tarkemmat käyttöehdot voidaan määritellä esimerkiksi erillisen lisenssin avulla. Lisenssillä viitataan käyttöön oikeuttavaan sopimukseen, jolla aineisto on julkaistu. Fiorettin (2011) mukaan avoimen datan asianmukainen lisensointi on välttämätöntä. Mikäli käyttäjällä ei esimerkiksi ole takeita siitä että aineistoa voidaan käyttää ilman rajoitteita, ei datalle juuri löydy hyödyntäjiä. Salliva lisensointi voidaan toteuttaa esimerkiksi Creative Commons tai Open Database -lisenssillä (Poikola et al. 2010, s. 36), jotka noudattavat avoimen datan periaatteita vapaasta datan käytöstä, muokkaamisesta ja jakamisesta. Lisenssin sisällön kannalta tärkeimpiä tekijöitä ovat yksinkertaisuus ja ymmärrettävyys, vapaa käyttöoikeus, oikeudenmukaisuus ja lisenssin läpinäkyvyys (Halonen 2012, s. 45). Lisenssi olisi hyvä jakaa avoimessa ja koneluettavassa muodossa, sillä koneluettavuus helpottaa aineistojen löydettävyyttä sekä parantaa yhdistelmäpalveluiden kehittäjien mahdollisuuksia seurata ehtojen noudattamista. (Poikola et al. 2010, s. 36)

2.3.4 Alkuperäisyys ja ajantasaisuus

Aineisto tulisi myös tarjota alkuperäisessä muodossaan ja alkuperäisellä tarkkuusasteella siten, että data julkaistaan mahdollisimman ajantasaisesti suhteutettuna aineiston sisällön muuttumisnopeuteen. (Poikola et al. 2010, s. 36) Esimerkiksi säätiedot muuttuvat huomattavasti useammin kuin kartta- ja paikkatiedot. Ubaldi (2013, s. 8) korostaa puolestaan ajantasaisuutta datan arvon näkökulmasta. Usein ajantasaisempi tieto on arvokkaampaa kuin historiatieto, sillä päätelmät ja analyysit datasta voidaan tällöin tehdä nopeammin ja määrittää tarvitut toimenpiteet.

Joissain tapauksissa yleistämällä ja laskemalla tarkkuustasoa voidaan muutoin yksityisyydensuojan kannalta arkaluontoinen aineisto saattaa avoimeksi (Poikola et al. 2010, s. 36). Tällaisia tapauksia voivat esimerkiksi olla arkaluonteisia tietoja sisältävät data-aineistot, joita ei ole sallittua julkaista alkuperäisessä muodossaan. Tällainen aineisto voidaan kuitenkin saattaa avoimeksi laskemalla tarkkuustasoa siten, etteivät esimerkiksi henkilötiedot ole pääteltävissä.

2.3.5 Laillinen ja vapaa uudelleenkäytettävyys

Borglundin & Engvallin (2014, s. 165) mukaan avoimen datan tärkeimpänä tavoitteena ei vain ole tuoda data kaikkien saataville, vaan myös uudelleenkäytettäväksi. Avoimen data-aineiston lisenssin ei siten tulisi rajoittaa datan uudelleen käyttöä tai jakelua (Gurin 2014a, s. 10), vaan käyttöehdoista tulisi myös käydä ilmi, että käyttäjällä on laillinen oikeus käyttää uudelleen aineistoa parhaaksi katsomallaan tavalla. Uudelleen jakelun tai käytön ehtona voi olla, että aineiston tekijät mainitaan. Ehdon noudattaminen ei saa kuitenkaan aiheuttaa kohtuuttomasti työtä ja onkin suotavaa, että aineiston yhteydessä toimitetaan lista niistä tahoista, jotka tulee mainita aineiston tekijöinä. (Open Knowledge Foundation 2015) Lisensointiehdot tulee olla esitetty selkeästi ja läpinäkyvästi siten, että ne kannustavat datan uudelleenkäyttöön (Poikola et al. 2010, s. 36).

Useat tutkimukset (Halonen 2012; Ubaldi 2013; Shadbolt & O'Hara 2013) ovat hyödyntäneet datan uudelleenkäytettävyyden mittarina Berners-Leen kehittämää niin sanottua viiden tähden mallia (Berners-Lee 2006), jonka tavoitteena on tuottaa avointa dataa, joka on mahdollisimman yhteensopivaa myös muiden avointen data-aineistojen kanssa. Mallin mukaan avoimen datan uudelleen käytettävyys paranee huomattavasti, kun se on mahdollista linkittää toisiin data-aineistoihin. (Ubaldi 2013, s. 26) Berners-Lee (2006) puhuu niin sanotusta linkitetystä avoimesta datasta.

Avoimen datan uudelleenkäytettävyyttä hankaloitetaan usein erillisellä rekisteröinnillä tai kulunvalvontamekanismilla, kuten salasanalla (Lindman et al. 2013). Usein perimmäisenä ajatuksena on datan tarjoajan mielenkiinto, mihin käyttötarkoitukseen dataa tullaan hyödyntämään. Poikolan et al. (2010, s. 36) tekemän kartoituksen mukaan useinkaan datan tarjoajalla ei ole varsinaisia tarpeita rajoittaa käyttötarkoituksia, vaan

datan uudelleenkäyttökohteet haluttiin tietää lähinnä oman toiminnan kehittämisen kannalta.

2.3.6 Maksuttomuus

Maksuttomuus viittaa puolestaan usein siihen, että data on saatavilla internetistä veloituksetta (Lindman et al. 2013) tai korkeintaan minimaalisella kustannuksella (Gurin 2014a, s. 10). Avointa dataa tulee myös voida hyödyntää vapaasti myös liiketoiminnassa kaupallisiin tarkoituksiin. Avointa dataa ei siis tulisi rajoittaa esimerkiksi vain tutkimus- ja opetuskäyttöön. Pienikin kustannus rajoittaa aineiston käyttöä merkittävästi siihen liittyvän vaivalloisuuden ja sopimusten vuoksi. (Poikola et al. 2010, s. 37)

Avoimen datan lisenssi ei myöskään saa määritellä, että hyödynnettävästä datasta tulisi maksaa rojalteja tai lisenssimaksuja, mikäli sitä käytetään kaupallisiin tarkoituksiin. Halonen (2012, s. 12-13) korostaakin, että dataa tulisi pitää yhteisenä resurssina, joka ei kärsi niukkuudesta. Dataa voidaan jakaa avoimesti ilman että sen arvo kärsii.

2.3.7 Koneluettavuus

Käyttäjälle on tärkeintä, että data on saatavilla ohjelmallisesti käsiteltävässä ja avoimessa esitysmuodossa, jotta datan käsittely ei edellytä minkään tietyn ohjelmiston käyttöä tai kaupallisen toimijan tuotteen hankkimista (Halonen 2012, s. 13). Kun data on saatavilla koneellisesti luettavassa muodossa, voidaan sitä käsitellä automaattisesti. (Manyika et al. 2013) Poikolan et al. (2010, s. 37) mukaan usein esimerkiksi organisaation verkkosivuilla tarjotaan aineistoja ei-koneluettavassa HTML-muodossa, jonka saattaminen koneluettavaan ja uudelleenkäytettävään muotoon saattaa olla hyvinkin työlästä. Dataformaatin avoimuus ja koneluettavuus ovat minimivaatimukset, jotta dataa päästään koneellisesti käsittelemään.

Datan koneluettavuus käsittää myös yleisemmin datan teknisen avoimuuden, eli standardit ja rajapinnat, jotka määrittelevät pääsyä käsiksi dataan. Rajapinta on tapa, jolla sovellus tai palvelu on yhteydessä dataan. (Lindman et al. 2013, s. 1241) Rajapintana käytetään yleisimmin ohjelmointirajapintaa, jonka avulla ohjelmat voivat tehdä pyyntöjä ja vaihtaa tietoja keskenään, sekä käyttää dataa oman sovelluksensa tarkoitukseen (Daintith & Wright 2008). Sen avulla käyttäjät voivat valita vain haluamiaan osia datasta ilman, että tarjolla oleva datatiedosto tulisi ladata kokonaisuudessaan. Ohjelmointirajapinta on tyypillisesti kytketty tietokantaan, jota päivitetään reaaliaikaisesti. Siten datan saanti rajapinnan kautta varmistaa, että data on myös ajantasaista. (Dietrich et al. 2015) Ohjelmointirajapinnan avulla hallinta siirtyy myös yli organisaatiorajojen, jolloin puhutaan niin sanotusta avoimesta arkkitehtuurista (Marton et al. 2013; Hjalmarsson et al. 2015). Toinen vaihtoehto datan jakamiselle on esimerkiksi FTP (File Transfer Protocol) -yhteyden käyttö, joka on alun perin suunniteltu tiedostojen siirtoon, mutta se ei ole enää niin suosittu vaihtoehto raakadatan jakamiselle. FTP -yhteydellä data jaetaan usein

kansiorakenteena, joka jättää vähemmän mahdollisuuksia datan räätälöinnille käyttäjän tarpeisiin. (Dietrich et al. 2015) Kuten Lindman et al. (2013, s.1243) toteavat, raakadata ei yleensä ole käytettävissä sellaisenaan, vaan sitä on jäsennettävä ja putsattava prosessin edetessä erilaisten toimintojen avulla. Datasta tulee esimerkiksi voida valita vain halutut attribuutit tai muokata data muuten käyttöön sopivaan muotoon.

2.3.8 Formaatin avoimuus

Data-aineisto tulisi myös pyrkiä avaamaan avoimessa formaatissa, jonka spesifikaatio on julkisesti ja vapaasti saatavilla, eikä se aseta rahallisia tai muita rajoitteita formaatin käytölle. (Poikola et al. 2010, s. 37) Mikäli data-aineistot ovat esimerkiksi ladattavissa verkosta Microsoft Excel -muodossa, voi Lindmanin et al. (2013, s. 1241) mukaan formaatin avoimuus olla kyseenalainen, sillä kyseessä on kaupallinen tiedostomuoto. Datan teknisen avoimuuden esteenä onkin usein pääsy dataan, mikäli pääsyä rajoitetaan dataformaatilla, joka ei ole kaikkien käytettävissä ilman erillistä ohjelmistoa. (Lindman et al. 2013)

Saatavilla oleva avoin data vaihtelee tällä hetkellä rakenteellisista tietokannoista tekstiin, taulukoihin, PDF -tiedostoihin ja moniin muihin. (Masip-Bruin et al. 2013, s. 332) Teknisesti avoin data tulisi kuitenkin jakaa muodossa, jota voidaan vaivattomasti käyttää uudelleen ja muokata. Näitä data formaatteja ovat esimerkiksi RDF, CSV ja XML (Bennett & Harvey 2009). Poikolan (2010, s. 37) mukaan aina formaatin avoimuuden takaamiseen ei kuitenkaan ole realistisia mahdollisuuksia, sillä esimerkiksi paikkatietojärjestelmien dataformaatit ovat usein valmistajakohtaisia, eikä avoimiin formaatteihin ole mahdollista siirtyä kuin vasta järjestelmäuudistuksen yhteydessä.

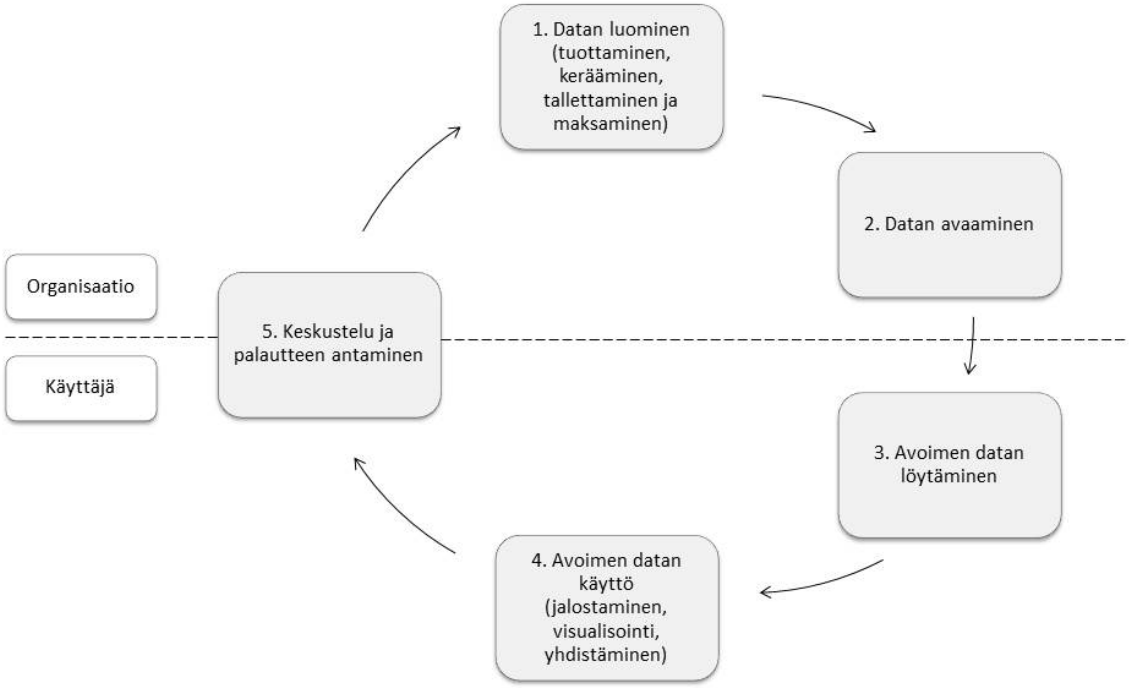
2.3.9 Ymmärrettävyys

Avoin data ja sen sisältö tulee myös kuvailla sekä dokumentoida kattavasti ja selkeästi esimerkiksi metadatan, käyttöesimerkkien ja laatumääritelmien avulla (Poikola et al. 2010, s. 38) Tällöin käyttäjä pystyy helposti omaksumaan, mistä datan sisällössä on kyse. Metadatalla tarkoitetaan tietoa tiedosta (Manyika et al. 2013), eli datan sisältöä kuvailevaa ja määrittelevää tietoa, joka tulisi tarjota avoimesti data-aineiston mukana. Metadata auttaa aineiston ymmärrettävyydessä ja sitä kautta parantaa myös merkittävästi datan uudelleenkäytettävyyttä. (Poikola et al. 2010, s. 37) Avoimen data-aineiston metadataan tulisi sisällyttää tiedot aineiston alkuperästä ja käyttöehdoissa. Metadata voi sisältää myös esimerkiksi käytettyjen termien ja lyhenteiden merkityksen kuvauksen.

2.4 Avoimen datan prosessimalli

Saatavilla oleva avoin data on raakadataa, jonka tulee käydä läpi useita eri vaiheita, jotta datasta tulee arvokasta. Halutaan raakadatasta sitten tuottaa palveluita tai sovelluksia,

analysoida sitä, yhdistää tai esittää se tietyllä tavalla, tarvitaan prosessi, jonka lopputuloksena datasta saadaan irti hyödyllistä informaatiota. (Lindman et al. 2013, s. 1241) Prosessi voidaan karkeasti jakaa viiteen perusaskeleeseen, joita ovat datan luominen ja kerääminen, datan avaaminen, avoimen datan löytäminen, käyttäminen sekä viimeisenä palautteen antaminen datan tarjoajalle. (Zuiderwijk et al. 2012) Avoimen datan prosessimalli sisältää sekä datan tuottajaorganisaation, että loppukäyttäjän näkökulman. Prosessi malli on esitetty kuvassa 4.



Kuva 4. *Avoimen datan prosessi (mukaillen Zuiderwijk et al. 2012)*

Avoimen datan prosessimalli lähtee luonnollisesti liikkeelle datan tuottajaorganisaatiossa datan luomisesta, joka sisältää datan tuottamisen, keräämisen, tallettamisen sekä mahdollisista kustannuksista huolehtimisen. Seuraavana on vuorossa itse datan avaaminen. Siihen lukeutuvat esimerkiksi datan muokkaaminen sopivaan muotoon ja dataformaatin sekä julkaisualustan valinta. Julkaisualustana voi toimia organisaation omat verkkosivut, kansallinen dataportaali tai muu alusta. (Zuiderwijk et al. 2012, s. 157) Kun data on avattu, tulee se saattaa yhtäläisesti kaikkien saataville ja avoimesti hyödynnettäväksi, jolloin prosessin kolmas vaihe voidaan suorittaa, eli käyttäjän on mahdollista löytää avoin data-aineisto ja päästä siihen helposti käsiksi. Neljäntenä vaiheena on avoimen datan käyttö, joka voi sisältää esimerkiksi datan jalostamista, yhdistämistä, parantamista tai visualisointia. Graves & Hendler (2013) toteavat tutkimuksessaan, että esimerkiksi visualisoinnin avulla voidaan helpottaa datan omaksumista ja päätelmien tekoa käyttäjien keskuudessa. Viimeisenä vaiheena on ideaalitilanteessa palautteen antaminen datan tarjonneelle organisaatiolle, jotta käyttäjältä saatuja tietoja voidaan käyttää edelleen parantamaan työprosesseja ja datan laatua. (Zuiderwijk et al. 2012, s. 157)

2.5 Datan avaamisen hyödyt

Euroopan komission (2011) tekemän selvityksen mukaan kyky käsitellä dataa älykkäästi on olennainen edellytys yhteiskunnallisiin haasteisiin vastaamiseksi. Datan avaamisella voidaan pyrkiä tehostamaan datasta saatavia hyötyjä sekä ymmärtämään dataa paremmin. Tällöin voidaan esimerkiksi onnistua parantamaan kansallisten terveydenhuoltojärjestelmien kestävyyttä, vastaamaan ympäristöhaasteisiin tai löytämään käyttämättömiä liiketoimintamahdollisuuksia. Useimmiten avoimen datan arvo muodostuu yhdistämällä avoimia datalähteitä organisaation sisäiseen, yksityisesti saatavilla olevaan dataan (Manyika et al. 2013, s. 7).

Avoimen datan tuottamia hyötyjä on tutkittu sekä datan tuottajan (Huijboom & Broek 2011) että loppukäyttäjän näkökulmasta (Janssen et al. 2012; Dietrich et al. 2015). Avoimesta datasta saatavat hyödyt voidaan jaotella Janssenin et al. (2012) tekemän tutkimuksen mukaan karkeasti kolmeen kategoriaan; poliittiset ja sosiaaliset, ekonomiset sekä toiminnalliset ja tekniset hyödyt. Empiirisesti toteutetun tutkimuksen mukaan haastateltavien ihmisten perusolettamuksena oli, että avoin data luo ja tuottaa enemmän lisäarvoa kuin näiden datajoukkojen myynti. Poliittisia ja sosiaalisia hyötyjä tuottavia tekijöitä pidettiin tutkimuksen mukaan tärkeimpinä. Taulukkoon 1 on kerätty esimerkkejä avoimen datan tuomista hyödyistä.

Taulukko 1. *Avoimen datan tuomia hyötyjä (mukaillen Janssen et al. 2012; Dietrich et al. 2015)*

Poliittiset ja sosiaaliset	• Toiminnan läpinäkyvyyden lisääntyminen • Parempi demokraattinen vastuullisuus • Kansalaisten osallistuttaminen • Vaikutusmahdollisuuksien lisääminen • Yhdenvertainen pääsy dataan • Palveluiden ja tyytyväisyyden parantaminen • Luottamuksen kasvu
Ekonomiset	• Talouskasvu • Kilpailukyvyn lisääntyminen • Kannustus uusiin innovaatioihin • Tietojen saatavuus sijoittajille ja yrityksille • Prosessien, tuotteiden ja palveluiden parantaminen
Toiminnalliset ja tekniset	• Päällekkäisyyksien poistaminen • Datan keräämisestä aiheutuvien kustannusten vähentyminen • Helpompi pääsy dataan • Datan ulkoinen laaduntarkastus • Julkisen ja yksityisen datan yhdistäminen

Suurimpina motivaatiotekijöinä datan avaamiselle ovat juuri tuotto-odotukset, joita toivotaan saatavan kun data avataan eri liiketoimille ja avoimille innovaatiolle hyödynnettäväksi. Datan avaaminen voi auttaa monimutkaisten ongelmien ratkaisussa, mutta myös antaa tukea päätöksentekoon ja osallistuttaa suurempia määriä ihmisiä tietojen analysointiin sekä lisätä läpinäkyvyyttä julkisten organisaatioiden toiminnassa ja poliittisessa päätöksenteossa. (Janssen et al. 2012) Datan avaaminen julkiseen käyttöön luo siten uusia markkinamahdollisuuksia aloitteleville start up -yrityksille ja muille pienille yrityksille. (Tammisto & Lindman 2012, s. 297) Avoimen datan odotetaan yleisimmin tuovan hyötyjä, kuten innovoinnin ja talouskasvun edistäminen. Kuitenkaan avoimen datan tuomia hyötyjä ei voida mitenkään tarkkaan määritellä ennalta (Gurin 2014a, s. 14), sillä datan avaamisella ei itsessään ole arvoa. Hyödyt muodostuvat vasta datan käytön seurauksena. (Janssen et al. 2012 s. 260) On siis mahdotonta arvioida, mikä yksittäinen hyötyjoukko nousee merkittävimmäksi. (Halonen 2012, s. 14)

2.6 Esteitä avoimen datan hyödyntämiselle

Avoimesta datasta puhuttaessa nousee usein esille määritelmä, jonka mukaan se on julkisesti kaikkien saatavilla, jolloin kaikilla on myös yhtäläinen mahdollisuus hyötyä siitä yhdistelemällä eri datalähteitä. (Ubaldin 2013; Dietrich et al. 2015) Gurstein (2011) esittää asiaan erilaisen näkökulman todetessaan, että kaikilla ei ole yhtäläistä pääsyä digitaaliseen infrastruktuuriin, tarvittaviin laitteistoihin tai ohjelmistoihin. Kaikilla ei myöskään ole mahdollisuutta taloudellisiin tai kasvatuksellisiin resursseihin ja taitoihin, joita tarvitaan, jotta avointa dataa voitaisiin hyödyntää tehokkaasti. Myös Graves & Hendler (2013, s. 136) toteavat tutkimuksessaan, että vaikka kaikilla olisi yhtäläinen pääsy dataan, ei se yksinään riitä hyötymään datasta. Teknisen asiantuntemuksen puute estää suurta osuutta väestöstä käyttämästä ja hyötymästä avoimesta datasta. Siten uudet löydökset avoimesta datasta ovat vain niiden käytettävissä, joilla on tieto tai taito hyötyä datasta.

Avoin data on oikeasti avointa vain siinä tapauksessa, mikäli se on helposti löydettävissä ja hyödynnettävissä. Tämä ei kuitenkaan Sushan et al. (2015, s. 182) mukaan useinkaan toteudu, vaan esteitä avoimen datan hyödyntämiselle on löydettävissä niin datan tarjoajan että käyttäjän näkökulmista (Zuiderwijk et al. 2012, s. 157). Datan tarjoajan näkökulmasta esteitä on tutkittu selkeästi enemmän (Huijboom & Broek 2011; Janssen et al. 2012; Zhang et al. 2005; Meijer & Thaens 2009; Conradie & Choenni 2014), kun taas loppukäyttäjän näkökulmasta esteet on usein sivuutettu (Blakemore & Craglia 2006; Janssen et al. 2012). Datan tarjoajan näkökulmasta dataa ei välttämättä haluta avata julkisesti kaikkien saataville ja puolestaan käyttäjän näkökulmasta saatavilla olevaa avointa dataa ei osata hyödyntää tarpeeksi yksinkertaisella tavalla. (Janssen et al. 2012, s. 261)

Esteet on jaoteltu Janssenin et al. (2012) tekemän tutkimuksen mukaan kuuteen kategoriaan; institutionaalisiin, monimutkaisuuteen, käyttöön ja osallistumiseen, lainsäädän-

töön, datan laatuun ja teknisiin tekijöihin liittyvät esteet. Institutionaaliset esteet koskevat lähes pelkästään datan tarjoajaa, kun taas monimutkaisuus sekä käyttö ja osallistuminen muodostuvat esteiksi useimmiten käyttäjän näkökulmasta. Muilla kategorioilla on yhtäläinen merkitys molemmille.

Taulukko 2. *Esteitä avoimen datan hyödyntämiselle (mukaillen Janssen et al. 2012; Huijboom & Broek 2011; Zuiderwijk et al. 2012)*

Institutionaaliset esteet	• Yhteisen toimintatavan puute datan avaamiseksi • Riskejä kaihtava organisaatiokulttuuri • Resurssien puute datan avaamiseksi • Yrityksen omat tuotto-odotukset datasta
Monimutkaisuus	• Kyvyttömyys löytää sopiva data • Datan puutteellinen saavutettavuus • Ei pääsyä alkuperäiseen dataan (vain jalostettuun) • Dataformaatit ja aineistot ovat liian monimutkaisia käsitellä ja käyttää helposti • Työvälineiden ja käyttötuen puute • Väärät johtopäätelmät datasta
Käyttö ja osallistuminen	• Ei aikaa tai kiinnostusta datan avaamiseksi • Ei kannustimia datan hyödyntämiselle • Datan ylikuorma • Datasta joudutaan maksamaan • Vaaditaan rekisteröinti ennen pääsyä dataan • Rikkomusten ja oikeusjuttujen uhka
Lainsäädäntö	• Yksityisyyden suojaa koskeva lainsäädäntö • Tekijänoikeus rikkomukset • Turvallisuuteen liittyvät uhat
Datan laatu	• Puutteet datassa • Vanhentunut tai virheellinen data • Tarvittavan datan puuttuminen kokonaan • Vain turvallinen data uskalletaan julkaista • Epäselvät arvot
Tekniset tekijät	• Hajanaisuus ohjelmistoissa ja dataformaateissa • Standardien puuttuminen • Puutteet metadatassa • Vanhat järjestelmät, jotka vaikeuttavat datan avaamista

Suurin osa esteistä on ratkaistavissa esimerkiksi toimintatapojen muutoksilla ja standardisoinnilla. Esteisiin tulisi puuttua vähintään julkishallinnollisten organisaatioiden ja valtioiden avoimen data politiikoissa ja strategioissa (Zuiderwijk et al. 2012), jotta estei-

tä pystyttäisiin poistamaan mahdollisimman tehokkaasti. Institutionaaliset esteet johtuvat suurimmaksi osaksi muutosten vastustamisesta. Dataa ei haluta avata omien tuotto-odotusten tai riskejä kaihtavan organisaatiokulttuurin vuoksi, sillä pelätään että datan avaaminen voi paljastaa epäkohtia organisaation toiminnasta tai pahimmillaan jopa johtaa syytteisiin. (Janssen et al. 2012 s. 261)

Käyttäjän näkökulmasta data saattaa olla liian monimutkaista hyödyntää omaan käyttötarkoitukseen, esimerkiksi dataformaattien eroavaisuuksista tai työkalujen ja ohjelmistojen puutteesta johtuen. Myös kynnys hyödyntää dataa kasvaa entisestään, mikäli siihen käsiksi pääsyä rajoitetaan esimerkiksi rekisteröinnillä tai käyttömaksulla. (Janssen et al. 2012) Avoin data on myös lähes poikkeuksetta maantieteellisesti rajoitettua, joka myös tarkoittaa sitä, että dataan perustuvat sovellukset ja palvelut ovat usein myös maantieteellisesti rajoitettuja (Hellberg & Hedström 2014, s. 47). Myöskään kaikkea mahdollista dataa ei edes voida tarjota avoimesti. Esimerkiksi Suomen henkilötietolaki (Finlex 2015) estää yksityisyyden suojaa loukkaavien tietojen jakamisen. Siten esimerkiksi potilastietoja tai muita arkaluonteisia tietoja ei voida avata siten, että henkilötiedot ovat saatavilla tai pääteltävissä, vaan data tulee anonymisoida.

Avoin data saattaa olla yksikertaisesti virheellistä tai puutteellista, mutta myös olennaisen metadatan puuttuminen saattaa johtaa vääriin tulkintoihin, mikäli esimerkiksi ei ole kerrottu aikaväliä, jolloin data on kerätty. (Janssen et al. 2012) Zuiderwijk et al. (2012) tekemässä tutkimuksessa puutteet metadatassa tuli esille kaikissa tutkituissa lähteissä. Tutkimuksen mukaan nykyisillään metadata on usein riittämätöntä tai sitä ei aina edes löydetä. Ilman metadataa datalta puuttuu selkeä semantiikka, joka selventää mitä data-attribuutit oikeasti kuvaavat. Usein data esitetäänkin tavalla, joka ei ole kaikille käyttäjille tarpeeksi ymmärrettävässä muodossa, jotta sen pohjalta voitaisiin tehdä perusteltuja päätelmiä. (Hoxha & Brahaj 2011) Teknisen osaamisen puute on merkittävä haaste sekä datan tarjoajan, että käyttäjän näkökulmasta. Esimerkiksi Daviesin (2013) tekemän maailman laajuisen kartoituksen mukaan julkisen liikenteen datastandardit ovat hyvin vakiintuneita, mutta vain 25% maailman maista on onnistunut avaamaan datan avoimesti koneellisesti luettavassa muodossa.

Esteet ovat usein myös yhteydessä toisiinsa, joka puolestaan lisää yleistä monimutkaisuutta. Tilanne johtuu pitkälti siitä, ettei esteitä ole juuri nostettu esille käyttäjän näkökulmasta eikä käyttäjän tarpeita ymmärretä. (Janssen et al. 2012, s. 266) Tämä selittää osaltaan tilannetta siitä, että vain rajattu osuus ihmisistä onnistuu oikeasti hyödyntämään avointa dataa tehokkaasti.

3. BIG DATA ANALYTIIKKA

Organisaatiot tuottavat biljoonia tavuja dataa esimerkiksi asiakkaista, toimittajista ja eri liiketoiminnoistaan. Digitaalista dataa syntyy kaikilla aloilla, jokaisessa taloudessa ja organisaatiossa, jotka jollain tavalla hyödyntävät digitaalista teknologiaa. (Chui et al. 2011, s. 1) Yleisesti on raportoitu, että noin 90% maailman datasta on luotu viimeisen kahden vuoden aikana (Davis & Patterson 2012, s. 2). Arvioiden mukaan kokonaisuudessaan datan määrä maailmassa kaksinkertaistuu aina 20 kuukauden välein. Vuoden 2015 loppupuolella pelkästään internetin dataliikenne oli noin zettatavun verran (Lausch et al. 2014, s. 5), joka vastaa biljoonaa gigatavua. Datasta on tullut tuotannon raaka-ainetta, joka tuottaa sekä taloudellista että yhteiskunnallista arvoa. (Tene & Polonetsky 2012, s. 63)

Pelkästään organisaatioiden liiketoiminnan sivutuotteena syntyy nykypäivänä jo valtava määrä digitaalista dataa. (Chui et al. 2011, s. 1) Asiakkaiden esimerkiksi vieraillessa yrityksen verkkosivustolla, tallentuu dataa selaamisesta, ostamisesta, jakamisesta, etsimisestä ja kommunikaatiosta. Esimerkiksi Walmartin palvelimet käsittelevät yli miljoona asiakastapahtumaa joka tunti ja tallentavat tietokantoihin yli 2,5 petatavua dataa. Myös Facebook käsittelee joka päivä lähes 500 teratavua käyttäjien lokitietoja sekä satoja teratavua kuvadataa, kun puolestaan eBay:n järjestelmät käsittelevät yli 100 petatavua dataa joka päivä. Toisaalta myös suihkulehtokone voi tuottaa jopa 10 teratavua toiminnallista dataa puolentunnin aikana. Se vastaa satoja teratavuja dataa jokaisesta Atlantin ylityksestä ja jos tämä luku kerrotaan noin 25 000 lennolla päivittäin, korostuu koneellisesti tuotetun datan suuruus entisestään. (Kambatla et al. 2014, s. 2562) Esimerkit luovat vain pienen katsauksen yhä lisääntyviin ja entisestään kasvaviin dataaineistoihin joita tuotetaan jo tänä päivänä. Koska teknologia kehittyy yhä edelleen, tulee myös datan määrä maailmassa mitä todennäköisemmin vain kasvamaan vuosien saatossa. (Maltby 2011, s. 1)

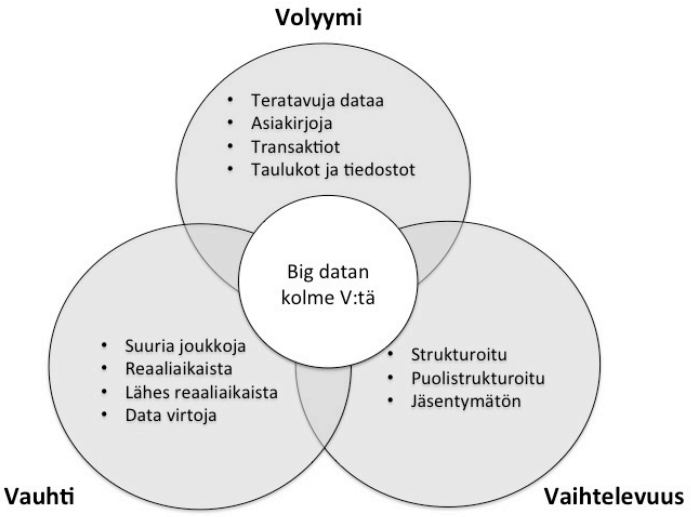
Datan määrän nopea kasvu johtuu pitkälti digitaalisten antureiden, viestinnän, laskennan ja varastoinnin kehityksestä. Tätä ilmiötä kuvaamaan Roger Magoulas kehitti termin big data. (Emani et al. 2014, s. 71) Viime vuosina big datasta onkin Maltbyn (2011, s. 1) mukaan tullut eräänlainen ilmiö liiketoiminnassa, informaatiotutkimuksessa, tietojärjestelmissä, tilastoiden käsittelyssä ja monilla muilla aloilla. Kyvystä varastoida ja yhdistellä dataa sekä hyödyntää tuloksia syvempien analyysien tukena on tullut jopa yhtä merkittävä trendi kuin Mooren laista tietojenkäsittelyssä, sen vastaavasta digitaalisesta varastoinnista, pilvilaskennan kustannusten alentumisesta sekä muista teknologisista esteistä. (Chui et al. 2011 s. 2)

3.1 Big datan määritelmä

Big datalla tarkoitetaan erityisen suurten ja järjestämättömien tietomassojen keräämistä, säilyttämistä ja ennen kaikkea analysointia tietoteknisten ratkaisujen avulla (Vakkuri 2013). Davisin & Pattersonin (2012, s. 44) ja Chuin et al. (2011, s. 1) mukaan big data sisältää liian suuren määrän dataa tallentaa, käsitellä ja analysoida perinteisillä tietokantaprotokollilla ja -ohjelmistoilla. Big data on avoimen datan tavoin hyvin uusi ilmiö (Gurin 2014b) ja seurausta datan määrän eksponentiaalisesta kasvusta viime vuosina. Jin et al. (2015, s. 59) lähestyvät big dataa makronäkökulmasta, jolloin big data voidaan nähdä siteenä, joka yhdistää ja integroi fyysisen maailman, ihmisyhteisön ja kyberavaruuden. Heidän mukaan big data voidaan jakaa kahteen luokkaan; dataan fyysisestä maailmasta, kuten sensoreista, tieteellisistä kokeista ja havainnoinneista sekä dataan ihmisyhteisöstä, jota usein hankitaan näiltä tahoilta, kuten terveys, rahoitus, talous ja internet.

Big data on usein peräisin passiivisista tietolähteistä ja se pidetään suurimmaksi osaksi organisaation sisäisenä tietona. (Gurin 2014a, s. 13) Organisaation sisäistä big dataa syntyy esimerkiksi erilaisista operatiivisista järjestelmistä, raporteista ja seurannasta. Maltbyn (2011, s. 2) mukaan big data ei yleiseen näkemyksen valossa viittaa pelkästään tietotulvan ongelmaan, vaan myös analyyttisiin menetelmiin, joiden avulla tietotulvaa voidaan hallita ja käyttää tuottavan ja käyttökelpoisen datan lähteenä.

Usein Big data määritellään niin sanotun kolmen V:n avulla, joita ovat vauhti, volyymi ja vaihtelevuus (Loshin 2013; Davis & Patterson 2012; Russom 2011; Krishnan 2013). Vauhdilla viitataan datan nopeaan lisääntymiseen, kun taas volyymi kuvastaa big datan suurta kokoa ja kykyä käsitellä sitä. Vaihtelevuudella puolestaan tarkoitetaan big datan sisältämiä datatyyppejä ja niiden erilaisia lähteitä. (Davis & Patterson 2012, s. 4) Kolmea V:tä ja niiden ominaisuuksia on havainnollistettu tarkemmin kuvassa 5.



Kuva 5. *Big datan kolme V:tä (mukaillen Russom 2011)*

Russomin (2011, s. 6-7) mukaan kooltaan big data määritellään usein teratavuina dataa, joka kuvastaa sen suurta volyymia. Dataa myös syntyy lisää hyvin nopeasti ja suurella vauhdilla, esimerkiksi reaaliaikaisista sensoreista. Koska big data koostuu usein monista eri lähteistä, sisältää se vaihtelevasti monia eri datatyyppejä. Sen vuoksi tarvitaan työkaluja jotka pystyvät käsittelemään sekä perinteisestä strukturoitua dataa, mutta myös puolistrukturoitua ja täysin jäsentämätöntä dataa, kuten tekstiä, kuvia ja videoita. (Jin et al. 2015, s. 59)

Jin et al. (2015) sekä Emani et al. (2015) laajentavat big datan määritelmää viiteen ominaisuuteen, lisäten edellä mainittuihin vielä monilaatuisuuden ja arvon realisoitumisen. Monilaatuisuus viittaa big datan epäjärjestelmällisyyteen ja luotettavuuden kyseenalaistamiseen, sillä datassa voi esiintyä epäselvyyksiä, päällekkäisyyksiä ja epäjohdonmukaisuuksia. Monilaatuisuudesta johtuen big datasta saatuja tuloksia ei aina voida esittää aukotta todeksi, mutta todennäköisyyksiä voidaan määrittää. Arvon muodostuminen puolestaan on yksi big datan tärkeimmistä ominaisuuksia ja se voidaan jakaa kahteen kategoriaan: analytiikan mahdollisuuksiin esimerkiksi päätöstenteon tukena ja uusien tarpeiden löytämisessä sekä uusien liiketoimintamallien mahdollisuuksiin uusien tuotteiden ja palveluiden kautta. (Emani et al. 2015, s. 72)

3.2 Edistyksellinen analytiikka ja big data

Kyky tehdä tarkkoja, oikea-aikaisia ja tehokkaita päätöksiä kaikilla organisaation päätöksenteon tasoilla, on tekijä, mikä yrityksen tänä päivänä usein erottaa kilpailuilla markkinoilla. Niin operatiivisella, taktisella kuin strategisellakin tasolla tulee pystyä käsittelemää asiakkaiden mieltymyksiä. (Bose 2009, s. 155) Yritykset ovatkin alkaneet hyödyntää edistyksellistä sekä ennakoivaa analytiikkaa analysoimaan erimuotoista dataa ja yhdistämään tietoja aikaisempien tietojen, nykyisten tapahtumien ja ennustettujen tulevaisuuden toimien välillä. Sisällyttämällä edistyksellisen analytiikan osaksi päivittäistä toimintaa, voivat organisaatiot tehdä päivittäisiä tietoon perustuvia päätöksiä liiketoimintatavoitteidensa saavuttamiseksi. (Apte et al. 2003)

Terminä edistyksellinen analytiikka tarkoittaa yksinkertaisesti edistyksellisen analytiikan teknologioiden hyödyntämistä vastaamaan datasta haettavien kysymysten tai ongelmien ratkaisuun. Edistyksellinen analytiikka ei siten itsessään ole teknologia, vaan ennemminkin joukko työkaluja, joita käytetään yhdessä datan analysointiin ja ennustamaan tuloksia ongelmien ratkaisuun. Perustana edistykselliselle analytiikalle toimivat datan yhdistely ja louhinta. Mitä enemmän dataa on kerätty ja yhdistelty, sitä enemmän voidaan havaita kaavoja ja suhteita data-aineistojen välillä. (Bose 2009, s.156) Franksin (2012, s. 187-188) mukaan edistyksellinen analytiikka menee ongelmien ratkaisussa ja tietoon perustuvien päätösten teossa selvästi pidemmälle kuin perinteinen data-analytiikka. Edistyksellinen analytiikka pyrkiikin selvittämään enemmän ja viemään päätelmiä pidemmälle, kuin pelkästään vastaamaan kysymyksiin mitä tapahtui ja millainen vaikutus tapahtuneella oli. Edistyksellinen analytiikka pyrkii lisäksi tunnistamaan

mikä tapahtuman aiheutti ja analysoimaan miten tapahtuneeseen voidaan varautua sekä mitä asialle voidaan tehdä tai miten siihen voidaan vaikuttaa tulevaisuudessa. Edistyksellinen analytiikka kattaa siten laajan kattauksen toimia, kuten ennakoivan analytiikan ja mallinnuksen, datan louhinnan, ennustamisen, optimoinnin sekä muita vastaavia toimia. Parantaakseen suorituskykyä edistyksellisen analytiikan avulla Bartonin & Courtin (2012) mukaan organisaatioiden tulisi kehittää vahvuuksiaan hyödyntää analytiikan tukena useita eri datalähteitä, niin organisaation sisäisiä kuin ulkoisiakin.

Big data analytiikka voidaan määritellä Russomin (2011, s. 8) mukaan edistyksellisten analytiikkatyökalujen käyttönä big datan jäsentämisessä. Big data analytiikan tavoitteena on joko kuvata menneisyyttä tai ennustaa tulevaisuutta (Vajjhalan et al. 2015, s. 490). Krishnan (2013, s. 251-252) määrittelee big data analytiikan tarkemmin perinteisten analytiikan ja datan louhinnan työkalujen yhdistelmäksi, jonka avulla voidaan käsitellä hyvin suurta määrää dataa ja esimerkiksi ennustaa asiakkaiden tai markkinoiden käyttäytymistä. Big data analytiikka nähdään siis työkaluna erityisen suurten datamäärien käsittelyyn, sillä sen avulla on mahdollista työstää suuria määriä sekä reaaliaikaista, että historiallista tietoa ja löytää uusia malleja tai poikkeavaisuuksia, jotka voivat osoittaa mahdollisuuksia uusille tuotteille ja palveluille tai tehokkaammille toimintatavoille. (Manyika et al. 2013, s. 1) Emanin et al. (2015, s. 71) mukaan big data analytiikan suurin viehätys onkin sen kyvyssä käsitellä hyvin suuria datamääriä.

Tänä päivänä yritykset hyödyntävät big dataa ja edistyksellistä analytiikkaa löytämään asioita, joita ei ennen ole havaittu. Big data analytiikan mahdollisuudet ovat erityisen tärkeitä juuri tällä hetkellä, kun viimeaikainen taloudellinen taantuma on pakottanut yrityksiä muuttumaan. Edistyksellisen analytiikan avulla yritykset voivat esimerkiksi ymmärtää liiketoiminnan nykytilaa ja seurata useita tekijöitä, kuten esimerkiksi asiakaskäyttäytymistä. (Russom 2011, s. 4) Kun ihmisten ja järjestelmien tuottaman datan määrä kasvaa eksponentiaalisesti, tietotulva on nähtävissä lähes kaikkialla. Big data analytiikka pyrkii hyödyntämään tietotulvan aiheuttamaa suurta data määrää ja käyttää sitä tuottavasti. (Maltby 2011, s. 6)

Big data ja analytiikka ovatkin nopeasti nousseet yritysjohtajien asialistalle, kun he ovat sivusta seuranneet miten Google, Amazon ja muut yritykset ovat onnistuneet syrjäyttämään kilpailijansa uusien, tehokkaiden ja erityisesti dataa hyödyntävien liiketoimintamallien avulla. (Barton & Court 2012) Maltblyn (2011, s. 1-2) tekemän tutkimuksen mukaan hyvin suuri osa tutkimuksista käyttää termejä big data ja big data analytiikka vaihdellen, tarkoittaen pääosin samaa asiaa. Havainto korostaakin yleistä mielipidettä siitä, että big data ei liity vaan tietotulvan tuomaan ongelmaan vaan viittaa myös edistyksellisiin analyysityökaluihin, joita käytetään organisaation datan hallinnassa ja kääntämään tietotulva ennemminkin osaksi tuottavaa ja hyödyllistä tietoa.

3.3 Hyödyntäminen liiketoiminnassa

Big data analytiikka tarjoaa perinteisille organisaatioille mahdollisuuden muuttaa toimintaansa innovatiivisempaan suuntaan, sillä sen avulla voidaan suunnitella, ennustaa, kasvattaa markkinoita ja palveluita kohti korkeampia tuloja. (Krishnan 2013, s. 252) Jinin et al. (2015, s. 60) mukaan big data aiheuttaa myös voimakkaan sysäyksen seuraavalle sukupolvelle informaatioteknologian alalla, viitaten pilvilaskentaan, mobiiliteknologiaan sekä sosiaaliseen liiketoimintaan.

Big data analytiikan hyödyt voidaan jaotella Ohlhorstin (2012) mukaan viiteen osa-alueeseen, joihin big data tuo lisäarvoa perinteisimpiin toimintatapoihin ja tekniikoihin verrattuna. Osa-alueita ovat perinteinen liiketoimintatiedon hallinta, tiedonlouhinta, tilastollinen analysointi, ennakoiva analytiikka sekä datan mallinnus. Big datan arvo koostuukin pääosin organisaatioiden loputtomasta pyrkimyksestä kohti kilpailuetua, joka kannustaa organisaatioita hyödyntämään suuria tietovarastoja ja ulkoisia datalähteitä paljastaakseen trendejä, tilastoja ja muita käytännöllisiä tietoja avuksi päättämään seuraavia siirtojaan markkinoilla. Big datan avulla voidaan parantaa analytiikkaa ja tuottaa dataa visualisointeja ja syvempiä analyyseja tarpeen mukaan. Big data analytiikan perimmäisenä tavoitteena on tuottaa uusia innovaatioita sekä ajaa muutosta avoimempaan sekä tietoisempaan suuntaan, jossa aikaansaannokset voidaan ennustaa ja niihin voidaan reagoida tarpeen mukaisella tavalla. (Krishnan 2013, s. 255)

Maltbyn (2011, s. 2) mukaan on olemassa lukemattomia analytiikan tekniikoita, joita voidaan hyödyntää big data hankkeissa. Se mitä tekniikoita analyysissä hyödynnetään, riippuu pitkälti datasta jota analysoidaan sekä tutkimuskysymyksistä, jotka halutaan ratkaista. Big data analytiikan mahdollistavat teknologiat kattavat monenlaisia matemaattisia, tilastollisia ja mallintamisen tekniikoita. (Kambatla et al. 2014)

Yleisimmät big data analytiikan työkalut ovat osa Apache -projektia ja rakennettu Hadoopin ympärille. (Emani et al. 2015, s. 73) Hadoop on avoimen lähdekoodin viitekehys, joka luo mahdollisuuden käsitellä suuria tietomääriä riippumatta datan rakenteesta. (Ishwarappa & Anuradha 2015, s. 321) Hadoop sisältää kaksi pääkomponenttia: Hadoop Distributed File System (HDFS) ja MapReduce. HDFS mahdollistaa korkean suorituskyvyn dataan pääsemiseksi ja MapdReduce luo puitteet suurten data-aineistojen rinnakkaiselle käsittelylle. (Chandarana & Vijayalakshmi 2014, s. 432) Hadoop on suunniteltu skaalautumaan yksittäisistä palvelimista jopa tuhansiin koneisiin tarjoten jokaiselle sekä paikallista laskentatehoa että varastointia. (Ishwarappa & Anuradha 2015, s. 321) Chandarana & Vijayalakshmi (2014) esittelevät Hadoopin lisäksi myös Project Storm sekä Apache Drill viitekehykset. Heidän mukaan Hadoop tarjoaa mahdollisuuden tallentaa ja prosessoida hyvin suuria määriä dataa, mutta ei sovellu tilanteisiin, jolloin ajantasaisuus on kriittinen tekijä. Project Storm mahdollistaa datavirtojen reaaliaikaisen analyysin, kun taas Apache Drill soveltuu parhaiten interaktiivisten ad-hoc -ratkaisujen toteuttamiseen.

Yhdysvalloissa big dataa hyödynnetään jo kaikilla liiketoimintasektoreilla, kun puolestaan Suomessa ja muualla Euroopassa ollaan Vakkurin (2013) tekemän selvityksen mukaan noin kolme vuotta Yhdysvaltoja jäljessä big data teknologioiden käyttöönotossa ja hyödyntämisessä. Työ- ja elinkeinoministeriön ICT 2015 työryhmän raportti (Ala-Pietilä & Pennanen 2013) tähdentääkin, että esimerkiksi Suomessa tulee aloittaa erillinen hanke big data osaamisen kehittämiseksi, jotta sitä osataan tulevaisuudessa hyödyntää tehokkaammin.

3.4 Avoimen datan ja big datan suhde

Yksi olennaisimmista big datan mahdollistamista hyödyistä on datojen yhdistely, johon avoin data on yksi parhaista raaka-aineista. (Peltola 2014) Avoin data tuo usein lisää syvyyttä big data analytiikalle sekä sen mahdollistamille uusille sovelluksilla ja palveluille. (Chui et al. 2013, s. 1) Avoin data ei Manyikan et al. (2013, s. 44) mukaan vain nopeuta big datan ja edistyksellisen analytiikan kasvua, vaan se voi myös luoda täysin uusia arvon lähteitä niin valmistajille, jälleenmyyjille kuin kuluttajillekin.

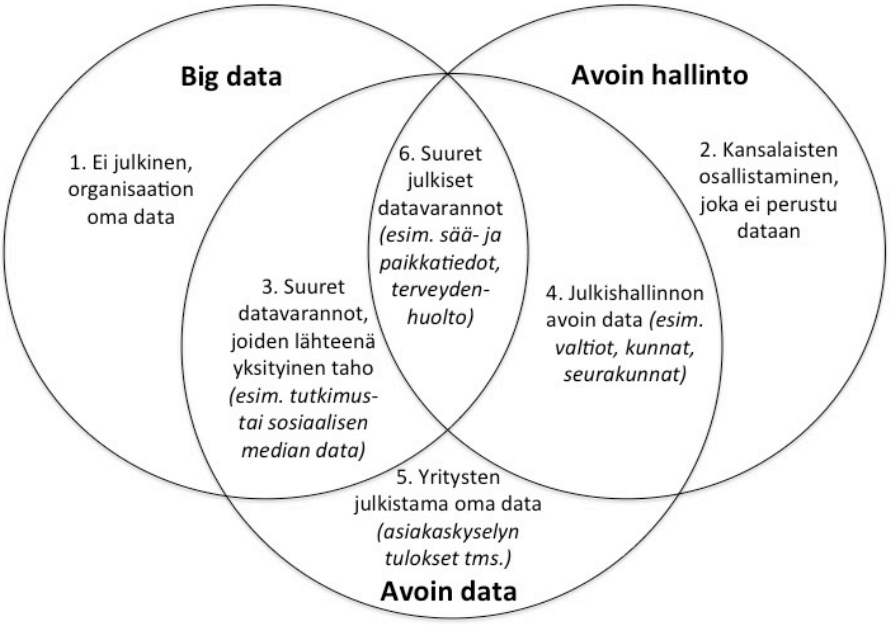
Avoimen datan ja big datan välistä suhdetta voidaankin kuvata seuraavasti; yleisesti ottaen big data sisältää avoimen datan ja avoin data on usein big dataa. Nykypäivänä big data ja avoin data ovat molemmat hyvin keskeisiä teemoja tietojärjestelmiin keskittyvässä tutkimuksessa. Useimmin big data määritellä osaksi liiketoiminta-analytiikkaa, mutta se usein myös muuttaa teknisiä sekä strategisia käytänteitä että tapoja, miten organisaatioiden tulisi hyödyntää dataa ja tehdä sen perusteella päätöksiä sekä parantaa suorituskykyään. (Marton et al. 2013, s. 3; Chen et al. 2012)

Vaikka big data ja avoin data ovat kaksi täysin erillistä käsitettä, voidaan niissä myös nähdä päällekkäisiä ja yhdistäviä teemoja. Yhdistämällä datatyyppien samankaltaisuuksia, voidaan ne koota yhteen käsitteeksi avoin big data. (Marton et al. 2013, s. 2) Avoimella big datalla viitataan siten hyvin suuriin, nopeasti muuttuviin ja sisällöltään vaihteleviin data-aineistoihin jotka on koottavissa avoimista datalähteistä. Myös Bedini et al. (2014) määrittelevät avoimen big datan hyvin suuriksi data-aineistoiksi, joita ovat tuottaneet esimerkiksi valtion virastot ja yritykset sekä edelleen julkaisseet aineistot avoimesti kaikkien saataville. He näkevät avoimen big datan erityisen mielenkiintoisena silloin, kun avointa dataa tarjoava portaali voidaan suoraan integroida big dataa käsittelevään alustaan tai ohjelmistoon, joka pystyy tarjoamaan hyödyllistä analytiikkaa kohtuullisessa ajassa.

Marton et al. (2013, s. 2) korostavat avoimen big data analytiikan tuomaa hyötyä erityisesti uusien löydösten tekemiseen data-aineistosta, joita ei muuten olisi havaittu. Heidän mukaansa tämä saavutetaan avoimen big datan kahden ulottuvuuden avulla, joita ovat järjestys ja relationaalisuus. Järjestys viittaa nykyaikaisiin teknologioihin, kuten hakukoneiden algoritmeihin, jotka mahdollistavat hyvin suurten datamäärien järjestämisen

kun data on avattu saataville. Relationaalisuus puolestaan viittaa data-aineistojen yhdisteltävyyteen.

Gurin (2014b) pohtii avoimen datan, big datan sekä avoimen hallinnon välistä suhdetta Venn-diagrammin avulla. Avoin hallinto nähdään kaaviossa julkishallinnon avoimen big datan tuottaja, mutta sekä big data, että avoin data voivat olla myös yksityisen organisaation luomaa. Kuvassa 6 on havainnollistettu kaavio esittää, miten big data ja avoin data liittyvät avoimen hallinnon laajaan käsitteeseen.



Kuva 6. *Avoimen datan ja big datan välinen suhde (mukaillen Gurin 2014b)*

Kaaviossa data on jaoteltu kuuteen eri kategoriaan. Big data, joka ei ole avointa dataa, voidaan nähdä yksityisellä sektorilla toimivan organisaation sisäisenä big datana. Hyvin suuri osa big datasta sijoittuu juuri tähän kategoriaan, sillä yritykset keräävät hyvin paljon dataa esimerkiksi tuotteistaan, asiakkaistaan ja heidän kulutustottumuksista. Toinen kategoria kuuluu puolestaan avoimen hallinnon käsitteen alle, jolla tarkoitetaan avoimen hallinnon käytänteitä. Käytänteet eivät perustu dataan, vaan toiminnan läpinäkyvyyteen ja kansalaisten osallistumiseen. Kolmas kategoria sijoittuu avoimen big datan alle, mutta ei kuitenkaan ole julkisten organisaatioiden tuottamaa. Tällaista dataa voi olla esimerkiksi sosiaalisen median avulla saatava data. Neljäntenä kategoriana nähdään julkishallinnon tuottama avoin data, joka sisältää esimerkiksi avoimesti saatavilla olevaa dataa valtioiden ja kuntien toiminnasta, tutkimuksista ja budjeteista. Viidentenä kategoriana määritellään yksityisten organisaatioiden tuottama avoin data, joka halutaan avata omien tarkoitusperien vuoksi. Dataa voidaan avata esimerkiksi parantamaan yrityksen mainetta tai tyydyttämään sijoittajien tarpeita. Kaikkien kolmen tahon keskelle sijoittuvana kategoriana nähdään avoin big data joka on usein julkistenorganisaatioiden tuottamaa, joka koostuu usein suurista avoimista datavarannoista, kuten sää- ja paikkatiedoista, terveydenhuollosta, liikenteestä tai budjetoinnista. (Gurin 2014b)

Martonin (2013, s. 5) mukaan erityisesti julkisten organisaatioiden tuottamat suuret avoimet data-aineistot voivat hyödyntää big datan tekniikoita ja teknologioita. Avoin big data nähdään potentiaalisena ulkoisena datalähteenä, joka yhdistettäessä organisaation sisäisiin datavarantoihin voi mahdollistaa täysin uudenlaisen ja entistä kattavamman analytiikan ja liiketoiminnan kehityksen. Ilmiöinä sekä avoin data että big data voivat muuttaa yritysten, julkisten hallinnon sekä koko yhteiskunnan toimintaa. Big data mahdollistaa ennennäkemättömän voiman ymmärtää, analysoida ja lopulta muuttaa maailmaa. Avoin data puolestaan takaa, että voima jaetaan ja maailmasta syntyy yhä oikeudenmukaisempi ja demokraattisempi. Erityisen voimakkaana voidaan Gurinin (2014b) mukaan pitää näiden kahden ilmiön yhdistelmää – avointa big dataa.

3.5 Avoin big data analytiikka

Tässä tutkimuksessa keskitytään tarkastelemaan avoimia big data-aineistoja, joita voidaan hyödyntää osana globaalin organisaation data-analytiikkaa ja löytää siten uusia havaintoja ja korrelaatioita eri datalähteiden väliltä. Tästä käytetään nimitystä avoin big data analytiikka. Edellä esitetyssä Gurinin (2014b) kaaviossa tutkimuksen aineisto sijoittuu keskiöön, eli pääosin hyvin suuriin julkisiin datavarantoihin. Viimeistään analysoitavasta aineistosta tulee big dataa, kun se yhdistetään muihin avoimiin data-aineistoihin tai organisaation sisäisiin ja tyypiltään vaihteleviin data-aineistoihin.

Jotta globaalissa organisaatiossa voidaan hyödyntää avointa big data analytiikkaa, tulee saatavilla olevaa avointa big dataa pystyä hyödyntämään ja käsittelemään tehokkaasti. Open Knowledge Foundation (2015) korostaa avoimen datan hyödynnettävyydessä yhteentoimivuutta muiden datalähteiden kanssa, jonka perusteella myös tässä tutkimuksessa määritellyt hyödynnettävyyden mittarit on asetettu. Avoimen datan hyödynnettävyyden mittareiksi asetettiin kymmenen erillistä arviointikriteeriä; saatavuus, kokonaisuus ja laatu, dataformaatti, arkkitehtuuri ja rajapinnat, käyttöehdot, kustannukset, metadatan kuvaus, maantieteellinen alue, havaintotarkkuus ja ajanjakso.

Vaikka tutkimuksessa käytetty mittaristo on määritelty pääosin hankkeen tilaajan kanssa yhteistyössä, nousevat samat kriteerit esille selkeästi myös kirjallisuudesta. Tutkimukset ovat aikaisemmin osoittaneet, että avoimen datan hyödyntämisen haasteille ja esteille on olemassa vähintään kolme pääkategoriaa, joita tulee tarkastella. (Zuiderwijk et al. 2012; Jansen et al. 2012; Zuiderwijk et al. 2014) Ensimmäiseksi haasteeksi on tunnistettu dataan käsiksi pääsyyn liittyvät haasteet, kuten datan luomiseen, avaamiseen, löytämiseen ja saatavuuteen liittyvät tekijät. Toiseksi datan käyttöä haittaavat tekijät, kuten käyttöehtojen rajoittaminen. Viimeisenä kategoriana nähdään tallennusta ja uudelleenkäyttöä haittaavat tekijät, kuten vaikeus keskustella tai tarkentaa aineiston sisältöön liittyviä epävarmuustekijöitä sekä palautteen antaminen datan tarjoajalle. (Zuiderwijk et al. 2012, s. 160)

Myös tässä tutkimuksessa avoimen datan hyödynnettävyyden haasteita kartoittavan mittariston arviointikriteerit kartoittavat kaikkia edellä esitettyjä kolmea kategoriaa ja niiden tunnuspiirteitä. Datan käsiksi pääsyä kartoitetaan saatavuuden ja kustannusten määrittämisen avulla. Puolestaan käyttöä haittaavia tekijöitä tarkasteltiin kokonaisuuden ja laadun, käyttöehtojen sekä maantieteellisen alueen, havaintotarkkuuden sekä ajanjakson tarkastelun kautta. Tallennusta ja uudelleenkäyttöä kartoitettiin puolestaan dataformaatin, arkkitehtuurin ja rajapintojen sekä metadatan kuvaamisen avulla. Tutkimuksessa käytettävät hyödynnettävyyden mittarit ja niiden esiintyvyys kirjallisuudessa on koottu taulukkoon 3.

Taulukko 3. *Avoimien data-aineistojen analysointi ja arviointikriteereiden esiintyminen kirjallisuudessa*

Saatavuus	Poikola et al. (2010), Open Knowledge Foundation (2015), Ubaldi (2013), Zuiderwijk et al. (2012)
Kokonaisuus ja laatu	Poikola et al. (2010), Ubaldi (2013), Zuiderwijk et al. (2012); Zuiderwijk et al. (2014)
Dataformaatti	Poikola et al. (2010), Ubaldi (2013), Berners-Lee (2006), Open Knowledge Foundation (2015), Zuiderwijk et al. (2012)
Arkkitehtuuri ja rajapinnat	Lindman et al. (2013), Poikola et al. (2010), Open Knowledge Foundation (2015); Zuiderwijk et al. (2014)
Käyttöehdot	Lindman et al. (2013), Poikola et al. (2010), Ubaldi (2013), Berners-Lee (2006), Open Knowledge Foundation (2015), Borglund & Engvall (2014), Zuiderwijk et al. (2012)
Kustannukset	Poikola et al. (2010), Open Knowledge Foundation (2015), Lindman et al. (2013), Zuiderwijk et al. (2012)
Metadatan kuvaus	Poikola et al. (2010), Zuiderwijk et al. (2012); Zuiderwijk et al. (2014)
Maantieteellinen alue	Hellberg & Hedström (2014)
Havaintotarkkuus	Janssen et al. (2012), Zuiderwijk et al. (2012)
Ajanjakso	Ubaldi (2013), Poikola et al. (2010)

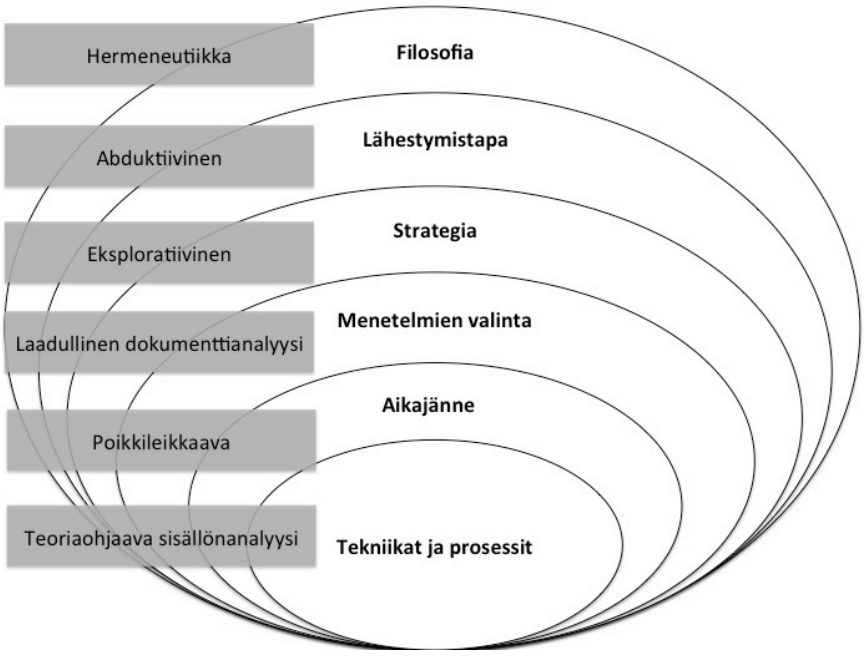
Saatavuus viittaa aineiston löydettävyyteen ja helppoon saavutettavuuteen, jolloin käyttäjä voi hyödyntää aineistoa milloin tahansa haluamallaan tavalla (Open Knowledge Foundation 2015). Kokonaisuus ja laatu kuvaa aineiston sisällön eheyttä (Dietrich et al. 2015) sekä sen laadukkuutta. Myös aineiston tekninen avoimuus, eli avoin dataformaatti, arkkitehtuuri sekä rajapinnat tulee huomioida, jolloin aineistoa pitää pystyä käsittelemään koneellisesti (Ubaldi 2013, s. 8) ja ilman tietyn kaupallisen toimijan tuotteen hankkimista (Lindman et al. 2013, s. 1241).

Käyttöehtojen määrittäminen on yksi avoimen datan olennaisimmista asioista, jota myös puoltavat taulukkoon listatut useat tutkimukset. Avoimen datan tuoma arvo korostuu nimenomaan silloin, kun data tuodaan kaikkien vapaasti uudelleenkäytettäväksi (Borglundi & Engvalli 2014, s. 165). Siten aineiston lisenssin ei tulisi rajoittaa mitenkään datan uudelleenkäyttöä tai edelleen jakelua (Gurin 2014a, s. 10), vaan ennemminkin korostaa datan vapaata hyödyntämismahdollisuutta. Kustannusten vaikutusta yhtenä mittarina voidaan puolestaan perustella sillä, että Poikolan et al. (2010, s. 37) mukaan pienikin kustannus rajoittaa merkittävästi aineiston käyttöä vaivalloisuuden ja sopimusten vuoksi. Sen vuoksi myös tutkimuksen taustalla olevassa hankkeessa keskityttiin lähes yksinomaan data-aineistoihin, jotka olivat maksutta saatavilla internetistä. Metadatan kuvaaminen puolestaan on tärkeää aineiston ymmärrettävyyden kannalta ja parantaa sitä kautta myös merkittävästi mahdollisuutta datan uudelleenkäyttöön (Poikola et al. 2010, s. 37).

Taulukossa esitetyistä mittareista maantieteellinen alue, havaintotarkkuus ja ajanjakso ovat kriteereitä, jotka on huomioitu erityisesti tarkasteltaessa avoimen data-aineiston hyödynnettävyyttä globaalin organisaation toiminnassa. Tällöin aineiston tulee kattaa maantieteellisesti mahdollisimman laaja otanta, joka useimmiten tarkoittaa monien eri aineistojen yhdistämistä. Tällöin myös havaintotarkkuuden ja ajanjakson tulee olla sama, jotta data-aineistot ovat vertailukelpoisia keskenään. Janssen et al. (2012) korostaa data-aineistojen ajanjakson kuvaamista jatkokäytön hyödyntämisen näkökulmasta, kun taas Hellberg & Hedström (2014) tarkastelevat maantieteellisen rajoittuvuuden haasteita, jolloin myös datan avulla tehtävät palvelut ja analyysit rajoittuvat vain tietylle alueelle. Sekä Ubaldi (2013), että Poikola et al. (2010) korostavat ajanjakson tarkastelemisessa myös datan ajantasaisuutta. Siksi ajanjakson tarkastelussa on otettu myös huomioon miten usein uutta dataa päivitetään, vai onko saatavilla pelkästään historiatietoon perustuvaa dataa. Tämän lisäksi on kartoitettu miten laajan aikavälin aineisto käsittää. Sekä maantieteellinen alue, aineiston havaintotarkkuus sekä ajanjakso ovat usein kuvattu osana aineiston metadataa, mutta tässä tutkimuksessa ne on arvioitu omina mittareina. Näiden kriteerien määrittelyyn on haluttu kiinnittää erityistä huomiota, sillä ne vaikuttavat suuressa määrin eri aineistojen yhdisteltävyyteen ja big data analytiikan mahdollisuuksiin.

4. TUTKIMUSMENETELMÄT JA AINEISTO

Tässä luvussa kuvataan tutkimuksen metodologisia valintoja sekä työssä käytettäviä tiedonkeruu- ja tutkimusmenetelmiä. Tutkimusmetodologialla tarkoitetaan sitä, miten tutkimus tulisi toteuttaa. Se käsittää teoreettiset ja filosofiset olettamukset, joihin tutkimus perustuu sekä näiden vaikutukset käytettävissä tutkimusmenetelmissä. (Saunders et al. 2009, s. 3) Näillä valinnoilla pyritään selkeyttämään tutkimuksen luonnetta ja lähestymistapaa määriteltyyn tutkimusongelmaan. Tutkimuksen tieteellisenä viitekehyksenä käytetään Saundersin et al. (2009) sipulimallia, joka on esitetty kuvassa 7. Kuvassa on esitetty myös tutkimuksessa tehdyt metodologiset valinnat, jotka perustellaan tarkemmin alaluvuissa.



Kuva 7. *Tutkimuksen tieteellinen viitekehys (mukaillen Saunders et al. 2009)*

Tutkimuksen viitekehys luo järjestyksessä katsauksen tutkimuksen tieteenfilosofiaan, lähestymistapaan, strategiaan, menetelmien valintaan, aikajänteeseen sekä lopuksi tekniikoihin ja prosesseihin. (Saunders et al. 2009, s.108) Viitekehyksen perustan luo tieteenfilosofia, edeten kohti tutkimuksen tarkempia ja yksityiskohtaisempia valintoja. Viitekehyksen viimeinen taso sisältää tarkemman katsauksen käytettyihin tekniikoihin ja prosesseihin, joilla tutkimusaineisto on valittu, kerätty sekä analysoitu.

4.1 Taustafilosofia ja tieteenkäsitys

Tieteenfilosofia käsittää tieteen tavoitteiden, menetelmien sekä tieteellisen tiedon tutkimisen filosofian näkökulmasta. Se tarkastelee lähinnä käsitteen- ja teorianmuodostusta, päättelyä ja selittämistä. (Olkkonen 1994, s. 15) Niiniluodon (1980, s. 21) tulkinnan mukaan tieteenfilosofialla tarkoitetaan filosofisen metodin soveltamista tieteeseen – tutkimustoimintaan ja sen tuloksiin. Molemmat näkemykset korostavat tutkimuksen taustalla toimivan filosofisen suuntauksen merkitystä erityisesti tutkimusmenetelmien sekä niiden avulla saatujen tulosten tulkinnassa. Eriaikoina vallinneiden käsitysten, tiedettä tutkineiden filosofien sekä eri tieteenalojen tavoitteiden ja perinteiden pohjalta on syntynyt erilaisia tieteenkäsityksiä (Olkkonen 1994, s. 26).

Olkkosen (1994, s. 26) mukaan merkittävimpiä tieteen taustalla olevia tieteenkäsitystä ovat hermeneutiikka ja positivismi. Hermeneutiikka voidaan kuvata tieteenkäsityksenä, joka korostaa tulkinnan, merkityksen, historian ja ymmärtämisen käsitteitä. Hermeneuttisessa tieteenkäsityksessä teoriataustaa ei välttämättä ole ja tutkijan subjektiivisuus on lähes välttämätöntä. Positivismi on puolestaan lähes päinvastainen tieteenkäsitys, joka nojaa yksinomaan todettuihin tosiasioihin hylkäämällä kaikki epävarmat asiat, jotka eivät ole toistettavissa. Keskeistä on siis tutkimuksen riippumattomuus tutkijasta. Positivismissa ongelma on myös strukturoitavissa ja lähdeaineistona pidetään aiempaa teoreettista tietoa. Hermeneutiikan takana on nähtävissä yleisempi filosofinen suuntaus idealismi ja positivismin takana realismi. Tutkimuksen tieteenkäsitys riippuu siis tutkijan tavasta tarkastella kohdetta (Saunders et al. 2009, s. 108).

Tässä tutkimuksessa hyödynnetään hermeneuttista tieteenkäsitystä, sillä tutkimustuloksissa korostuu subjektiivinen tulkinta eikä tutkimus pohjaudu vahvaa teoriataustaan. Hermeneuttinen tieteenkäsitys on valittu myös sen soveltuvuudesta tilanteisiin, joissa tutkittavaa ilmiötä ja sen merkitystä pyritään ymmärtämään tarkemmin. Tieteenkäsityksen valintaa perustelee myös Olkkosen (1994, s. 37) korostama hermeneuttisen tieteenkäsityksen sopivuus tilanteisiin, joissa ilmiö on uusi, tapauksia on vain vähän tai ne ovat vaikeasti strukturoitavissa. Uusi ilmiö pyritään siis kuvaamaan mahdollisimman tarkasti tietyssä tapauksessa ja sille etsitään mahdollisia selityksiä.

4.2 Lähestymistapa

Tutkimuksen lähestymistapa selittää tutkimuksen suhdetta teoriaan. Lähestymistapana käytetään perinteisesti joko deduktiivista tai induktiivista lähestymistä (Saunders et al. 2009, s. 106). Deduktiivinen päättely korostaa erikoisempien väitteiden johtamista yleisistä totuuksista ja se esiintyy useimmiten teoreettisessa tutkimuksessa. Induktiivinen päättely puolestaan yleistää väitteen johtamisella erikoisista tunnetuista tosiasioista ja se on tyypillistä empiiriselle tutkimukselle, jossa yksittäistapausten joukosta tilastollisesti päätellään koko populaatiota koskevia ominaisuuksia ja ilmiöitä. (Olkkonen 1994, s. 29-

30) Deduktiivisen ja induktiivisen tutkimuksen tunnuspiirteitä on esitetty tarkemmin taulukossa 4.

Taulukko 4. *Deduktiivisen ja induktiivisen tutkimuksen eroja (mukaillen Saunders et al. 2009)*

Deduktiivinen lähestymistapa	Induktiivinen lähestymistapa
• Tieteelliset periaatteet • Siirrytään teoriasta tietoon • Määrällisen aineiston kerääminen • Tutkijan riippumattomuus tutkimustuloksiin • Tarkkaan jäsennelty lähestymistapa ja rakenne	• Läheinen ymmärrys tutkimuksen kontekstista • Laadullisen aineiston kerääminen • Ihmisten ja tapahtumien liittäminen aiheeseen • Tutkija osana tutkimusprosessia • Rakenne joustavampi muutoksille

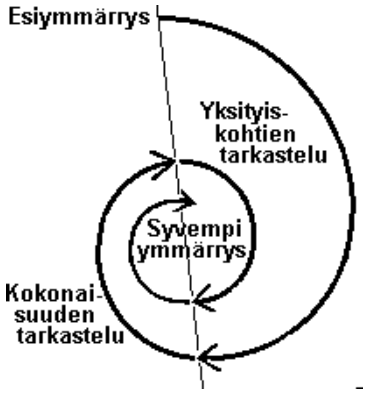
Deduktiivisen ja induktiivisen lähestymistavan lisäksi Tuomi & Sarajärvi (2003, s. 95-97) esittelevät abduktiivisen lähestymistavan, jonka mukaan teorian muodostus on mahdollista silloin kun havaintojen tekoon liittyy jokin johtoajatus. Abduktiivisesta tutkimuksesta puhutaan usein teoriaohjaavana päättelynä, sillä tutkijan ajatteluprosessissa vaihtelevat sekä aineistolähtöisyys että valmiit mallit. Abduktiivinen päättely voidaan siten nähdä olevan deduktiivisen ja induktiivisen päättelyn väliltä, sisältäen piirteitä molemmista. Tässä tutkimuksessa käytetään abduktiivista lähestymistapaa, sillä tutkimus perustuu laadulliseen aineistoon, jota on analysoitu teoriaan pohjautuvilla menetelmillä, mutta analyysi ei suoraa pohjaa teoriaan vaan se toimii johtoajatuksena tutkimuksen teossa.

4.3 Tutkimusstrategia

Tutkimusstrategialla tarkoitetaan tutkimuksen menetelmällisten ratkaisujen kokonaisuutta. (Hirsjärvi et al. 2007, s. 128) Saundersin et al. (2009, s. 141) mukaan tutkimusstrategian tehtävänä on ohjata tutkimusongelman asettelua ja tavoitteiden asettamista. Kirjallisuudessa tutkimusstrategioita tyypitellään monin eri tavoin, joista perinteisimpiä tutkimusstrategioita ovat kokeellinen tutkimus, survey-tutkimus sekä tapaustutkimus. Kokeellisessa tutkimuksessa mitataan yhden käsiteltävän muuttujan vaikutusta toiseen muuttujaan, kun survey-tutkimus puolestaan kerää tietoa standardoidussa muodossa joukolta ihmisiä, esimerkiksi kyselylomakkeen avulla. Tapaustutkimuksessa kerätään yksityiskohtaista tietoa yksittäisestä tapauksesta tai pienestä joukosta toisiinsa suhteessa olevia tapauksia. (Hirsjärvi et al. 2007, s. 130)

Tässä tutkimuksessa ei voida tehokkaasti hyödyntää perinteisiä tutkimusstrategioita, sillä tutkimus kohdistuu hyvin uudelle tutkimusalueelle, eikä tutkimuksen hypoteesista ollut selkeää käsitystä etukäteen. Tutkimus on siten luonteeltaan eksploratiiviinen, eli

uutta tietoa kartoittava tutkimus. Eksploratiivinen tutkimus selvittää vähemmän tunnettuja ilmiöitä tai se voi löytää kokonaan uusia ilmiöitä tai näkökulmia ja se voi myös kehittää hypoteeseja. (Tuomi 2007, s. 126) Saundersin et al. (2009, s. 139) mukaan eksploratiivinen tutkimus on erityisen hyödyllinen, kun tutkimusongelman luonne ei ole yksiselitteinen ja ymmärrystä halutaan selventää. Strategian suurimpana etuna joustavuus ja mukautuvuus tutkimuksen kulun mukaisesti, mikäli uuden havainnon tai oivalluksen seurauksena tutkimuksen suuntaa tulee muuttaa. Eksploratiivinen tutkimus ei etene useinkaan lineaarisesti, vaan syklisesti hermeneuttisessa kehässä, johon kuuluu tutkijan jatkuva oppiminen. Hermeneuttista kehää ja sen vaiheita on havainnollistettu kuvassa 8.



Kuva 8. *Hermeneuttinen kehä (Routio 1990)*

Hermeneuttinen kehä alkaa aiheen esiymmärryksestä ja etenee kohti syvempää ymmärrystä. Hermeneuttisen kehän ajatus korostaa sitä, että tutkimuksen tarkoituksena ei ole palata alkupisteeseen vaan edetä esiymmärryksestä syvemmälle. (Routio 1990) Kehällä tutkija käy läpi aineistoa useaan kertaa yrittäen vapautua omista esteistään ymmärtää tutkimuskohdetta. Kehää kiertäessään tutkija pääsee koko ajan lähemmäksi tutkimuskohdettaan ja toisaalta syventää itseymmärrystään. (Anttila 1998) Tässä tutkimuksessa teoriapohja on antanut esiymmärryksen ja johtoajatuksen tutkimuksen toteuttamiselle ja tutkimuksen edetessä aihepiiristä on muodostunut syvempi ymmärrys, kun työn empiirisiä tuloksia on tarkasteltu ja analysoitu.

4.4 Tutkimusmenetelmä

Perinteisesti tutkimusmenetelmät on jaoteltu kahteen osaan: kvantitatiiviseen, eli määrälliseen tutkimukseen sekä kvalitatiiviseen, eli laadulliseen tutkimukseen. (Hirsjärvi et al. 2007, s. 131-133; Tuomi 2007, s. 94) Alasuutarin (1999, s. 26) mukaan selkeää kahtiajakoa tutkimusmenetelmien välillä ei kuitenkaan aina voida tehdä, sillä useissa tutkimuksissa on piirteitä molemmista. Tällaista tutkimusmenetelmää kutsutaan monimenetelmäisyydeksi, kun tutkimuksessa yhdistetään sekä kvalitatiivisen, että kvantitatiivisen tutkimuksen menetelmiä. (Saunders et al. 2009, s. 108)

Tässä työssä tutkimusongelma rakentuu puhtaasti tutkimuskohteen ymmärtämisen ja haasteiden kartoittamisen ympärille, eli puhutaan laadullisesta tutkimuksesta. Hirsijärven et al. (2007, s. 177) mukaan laadullisessa tutkimuksessa yksittäisiä tapauksia tarkastelemalla kyllin tarkasti saadaan näkyviin myös se, mikä ilmiössä on merkittävää ja mikä toistuu usein tarkasteltaessa ilmiötä yleisellä tasolla. Myös tässä tutkimuksessa kartoitetaan vain hyvin pieni otos saatavilla olevista avoimista datalähteistä, mutta otoksesta löydettävien tulosten ilmiöitä voidaan myös pyrkiä yleistämään, kun merkittävimmät tekijät nousevat esille useammasta lähteestä. Tuomen (2007, s. 97) mukaan on myös huomioitava, että laadullisessa tutkimuksessa tieto, jota kerätään, liittyy aina ihmisten tuottamiin merkityksiin ja se suosii aineistolähtöistä analyysiä. Laadullinen tutkimus voidaan siten nähdä kokonaisuutena, jossa aineiston keräämistä ja analyysiä ei voida puhtaasti erottaa toisistaan (Tuomi & Sarajärvi 2009, s. 68).

Tuomen (2007, s. 96) mukaan laadullisen tutkimuksen lajeja on nähtävissä useita erilaisia, joiden taustalla on monenkirjava joukko erilaisia perinteitä. Tässä tutkimuksessa hyödynnettävä aineisto koostuu kirjalliseen muotoon saatetusta materiaalista, joten tutkimusmenetelmänä käytetään laadullista dokumenttianalyysiä. Dokumenttianalyysillä tarkoitetaan kaiken sellaisen todennettavissa olevan tutkimusaineiston analyysia, jota ei saada kokoon suorien tai välittömien havaintojen teolla. Dokumenteilla tarkoitetaan laajasti ottaen kaikenlaista ilmiötä dokumentoivaa aineistoa, joka voi olla esimerkiksi julkaistuja tekstejä, arkistomateriaalia, kertomuksia, elämänkertoja, valokuvia, tai videonauhoitteita. (Anttila 1998) Tutkimuksen tavoitteena on kartoittaa avoimesti saatavilla olevia datalähteitä, jonka vuoksi on luonnollista hyödyntää valmiiksi avattuja dataaineistoja, eikä pyrkiä kokoamaan näitä itse. Anttilan (1998) mukaan valmiin aineiston käyttö on joskus ainoa mahdollisuus saada kootuksi tietoa jostakin tietystä aiheesta, kuten on myös tämän tutkimuksen tapauksessa.

4.5 Aikajänne

Tutkimuksen aikajänne voidaan määritellä Saundersin et al. (2009, s. 155) mukaan joko poikkileikkaavaksi tai pitkittäiseksi. Poikkileikkaavalla aikajänteellä tarkoitetaan tietyn ajan hetken tarkastelua ja sen hetkisen tilanteen kuvaamista. Puolestaan pitkittäinen aikajänne kattaa tietyn ajan mittaisen tarkastelujakson tarkastelun, joka voi esimerkiksi olla muutaman kuukauden tai useamman vuoden mittainen ja kuvaa siten tilanteen kehitystä ajansuhteen. Saunders et al. (2009, s. 155) havainnollistavat eri aikajänteitä esimerkillä, jossa poikkileikkaus on verrattavissa tietyllä ajanhetkellä otettuun valokuvaan, kun taas pitkittäinen ajanjakso tarkoittaa useasta kuvasta tehtyä sarjaa.

Tässä tutkimuksessa käytetään poikkileikkaavaa aikajännettä. Aikajänteen valintaa voidaan perustella sillä, että tutkimuksessa ei ole ajallista ulottuvuutta suhteessa tutkittavan aineiston seurantaan. Kuten Tuomi (2007, s. 123) määrittelee, tutkimus kuvailee millainen tilanne on tutkimuksen toteuttamisen ajankohtana.

4.6 Tekniikat ja prosessit

Seuraavaksi kuvataan tarkemmin tekniikat ja prosessit, joiden avulla tutkimusaineisto on valittu, kerätty sekä analysoitu. Tutkimusaineiston valintaan on vaikuttanut pitkälti tutkimuksen taustalla olevan hankkeen tuomat rajoitteet ja vaatimukset, jotka ovat lähtöisin hankkeen tilaajalta. Tutkimuksen analyysimenetelmäksi valittiin sisällönanalyysi, sillä tarkoituksena oli kartoittaa saatavilla olevien datalähteiden sisältöä ja hyödynnettävyyttä. Tutkimuksen aineiston keräämiseen ja analysointiin käytettyä prosessia, sekä sen sisältämiä hyödynnettävyyden mittareita, on myös esitetty luvussa tarkemmin.

4.6.1 Tutkimusaineiston valinta

Tutkimusaineisto koostuu avoimista datalähteistä, jotka sisältävät pääosin sää- ja paikkatietoa sekä tuuliturbiineista ja sähköverkosta saatavilla olevaa dataa. Tutkimuksen taustalla olevassa hankkeessa kartoitettiin tuulivoimaloiden tuuliturbiineihin vaikuttavia ulkoisia tekijöitä ja niistä löytyvää avointa dataa, jota voitaisiin myöhemmin hyödyntää organisaation big data analytiikassa. Tutkimusaineiston valintaan on siten vaikuttanut pitkälti hankkeen tilaajana toimineen organisaation vaatimukset, joiden mukaan tuuliturbiinien toimintaan vaikuttavaa avointa dataa haluttiin mahdollisimman laajalta maantieteelliseltä otannalta viimeisten 15 vuoden ajalta. Tärkeimpinä maantieteellisinä alueina olivat Eurooppa, Yhdysvallat, Kanada, Australia sekä Etelä-Amerikka, jossa tilaajaorganisaation tuuliturbiinit ovat runsaimmassa käytössä. Tutkimusaineisto koostuu 16 avoimesta datalähteestä, jotka on tarkemmin kuvattu liitteessä 1. Jokaisesta avoimesta datalähteestä on koottu erillinen taulukko, joka sisältää lyhyen kuvauksen kaikkien datalähteille asetettujen arviointikriteereiden mukaisista tekijöistä sekä tehdyistä huomioista. Tämän lisäksi liitteeseen on koottu lyhyt katsaus jokaisen tutkimusaineistoon kuuluvan avoimen data-aineiston sisällöstä. Tutkimusaineiston valintaan on vaikuttanut myös hankkeen tilaajaorganisaation ehdottamat potentiaaliset datalähteet, jotka organisaatio oli ennalta kartoittanut vaihtoehtoisiksi datalähteiksi. Kartoitusta on tämän jälkeen laajennettu myös muihin vielä tunnistamattomiin avoimiin datalähteisiin edellä esitettyjen vaatimusten mukaisesti.

Tutkimusaineiston valinta aloitettiin määrittämällä tuuliturbiinien toimintaan vaikuttavat tekijät, jotka edelleen jaettiin neljään eri kategoriaan: sää-, paikkatieto-, tuuliturbiini- sekä sähköverkkodataan. Tämän jälkeen käytiin läpi hankkeen tilaajan ehdottamat potentiaaliset datalähteet sekä alettiin kartoittaa muita avoimia datalähteitä selaamalla internetistä löytyviä portaaleita ja sivustoja. Kartoitus aloitettiin suurista julkisista dataportaaleista, kuten Yhdysvaltojen ja Euroopan avoimen datan portaaleista, jonka jälkeen siirryttiin tarkastelemaan muita verkosta saatavilla olevia avoimia datalähteitä. Datalähteiden etsimiseen käytettiin aihepiiriin liittyviä hakulausekkeita ja -sanoja, jotka määrittyivät tarkemmin tutkimuksen edetessä, kun tuntemus aihepiiristä ja käytettävästä sanastosta lisääntyi. Aineistojen kartoitus aloitettiin seuraavan hakulausekkeen avulla:

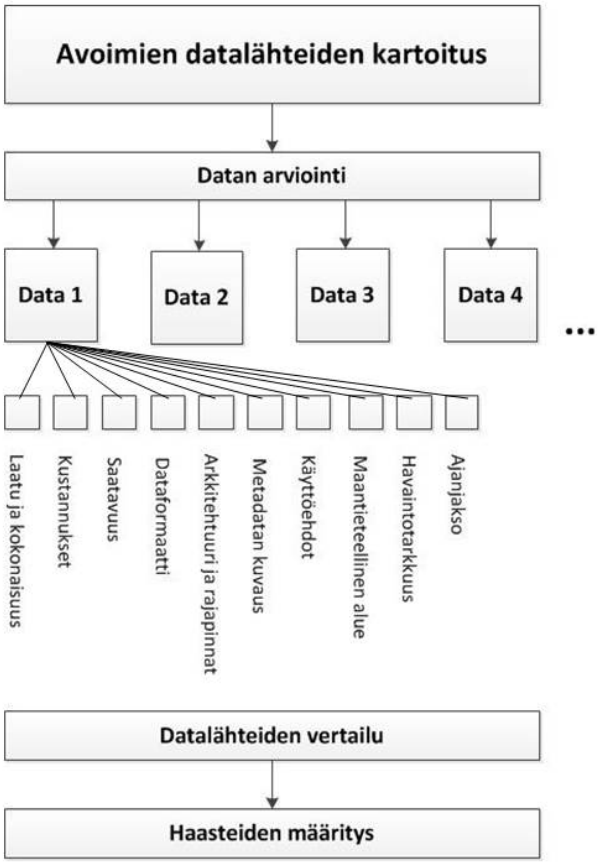
”open weather data OR open geographic data OR open electricity data OR open wind turbine data”. Myöhemmin hyödynnettiin myös hakusanoja ”open GIS data” sekä ”open data AND global weather”, jotka auttoivat kartoittamaan suuremman määrän avoimia datalähteitä, kun hakulausekkeisiin vielä lisättiin halutun maanosan tai valtion nimi, esimerkiksi ”open weather data finland”.

Tutkimusaineiston valintaan on vaikuttanut hankkeen tilaajaorganisaation vaatimusten ja ehdottamien datalähteiden lisäksi datan löydettävyys sekä tutkijan oma subjektiivinen tulkinta avoimen datalähteen sopivuudesta ja vaatimusten täyttymisestä. Tutkimusaineiston valinta on toteutettu käymällä läpi hyvin suuri määrä avointa dataa tarjoavia sivustoja ja portaaleita, joiden pohjalta on valittu aineistoksi sellaiset avoimet datalähteet, joissa toteutuu parhaiten tilaajaorganisaation vaatimukset datalle.

4.6.2 Analyysimenetelmä

Tutkimusmenetelmänä dokumenttianalyysi mahdollistaa sekä kvantitatiivisen, että kvalitatiivisen tutkimusaineiston analysoinnin. Kvantitatiivisesta analyysistä voidaan käyttää termiä sisällönerittely, joka tarkoittaa dokumenttien kuvaamista määrällisesti, esimerkiksi numeroin ja kvalitatiivisesta analyysistä puolestaan termiä sisällönanalyysi, joka tarkoittaa dokumenttien sanallista kuvaamista ja sen tavoitteena on löytää merkittävää sisältöä. (Ojasalo et al. 2014, s. 137) Laadulliseen aineistoon perustuen tutkimuksen analyysimenetelmänä käytetään sisällönanalyysiä, jolla voidaan analysoida dokumentteja systemaattisesti ja objektiivisesti. Sisällönanalyysilla pyritään järjestämään tutkimusaineisto selkeään ja tiiviiseen muotoon kadottamatta sen sisältämää informaatioita. (Tuomi & Sarajärvi 2009, s. 103-108)

Anttilan (1998) mukaan sisällönanalyysi on työväline, jolla voidaan tuottaa uutta tietoa, uusia näkemyksiä sekä saattaa esiin piileviä tosiasioita. Tarkemmin sisällönanalyysi voidaan tässä tutkimuksessa määritellä teoriaohjaavaksi sisällönanalyysiksi, joka etenee lähtökohdiltaan aineiston ehdoilla. Tällöin analyysi ei suoraan perustu teoriaan, mutta kytkennät siihen ovat havaittavissa. (Tuomi & Sarajärvi 2009, s. 109–116) Tässä tutkimuksessa aineistosta tehdyille löydöksille etsitään tulkintojen tueksi teoriasta selityksiä tai vahvistusta. Kuvassa 9 on havainnollistettu tarkemmin avoimien datalähteiden kartoitukseen ja analysointiin käytettyä prosessia.



Kuva 9. *Avoimien datalähteiden kartoitus ja analysointi*

Kuten Tuomi & Sarajärvi (2009, s. 108) määrittelevät, aineiston laadullinen käsittely perustuu loogiseen päättelyyn ja tulkintaan, jossa tutkimusaineisto ensin hajotetaan osiin, käsitteellistetään ja kootaan uudelleen loogiseksi kokonaisuudeksi. Avoimien datalähteiden kartoitukseen ja analysointiin käytetty prosessi käsitti neljä päävaihetta; datalähteiden kartoituksen, arvioinnin, vertailun sekä haasteiden määrittelyn. Teoriaohjaavassa sisällönanalyysissä on hyödynnetty erikseen määritettyjä kriteereitä, joilla avoimen datalähteen hyödynnettävyyttä on arvioitu.

Aineiston keräämisen ensimmäisenä vaiheena kartoitettiin saatavilla olevat avoimet datalähteet, jotka täyttivät kokonaan tai suurimmalta osin niille asetetut tavoitteet. Käytännössä aineiston kartoitus tehtiin käymällä läpi toimeksiantajan ennalta kartoittamat avoimet datalähteet ja käymällä läpi millaista avointa dataa on tarjolla. Tämän jälkeen kartoitusta laajennettiin muihin avoimiin datalähteisiin hyödyntämällä hakugeneraattoreita, eri hakusanoja sekä avoimen datan portaaleja. Erilaisia avointa dataa tarjoavia sivustoja käytiin läpi useita kymmeniä, joista valittiin 16 erillistä datalähdettä niiden tarjoaman avoimen datan perusteella.

Tämän jälkeen jokainen avoin datalähde arvioitiin ennalta määriteltyjen kriteerien avulla. Arviointi toteutettiin kartoittamalla avoimen datan hyödynnettävyyttä erilli-

sen arviointikriteeristön avulla, jossa tarkastellaan avoimen data-aineiston saatavuutta, kokonaisuutta ja laatua, dataformaatin avoimuutta ja koneluettavuutta, teknistä avoimuutta arkkitehtuurikuvausten ja rajapintojen avulla, käyttöehtoja ja uudelleenkäytettävyyden mahdollisuutta, kustannuksia, sekä datan ymmärrettävyyttä metadatan kuvaamisella. Tämän lisäksi hankkeen tilaajan rajoitteiden mukaisesti tarkasteltiin datan maantieteellistä kattavuutta, havaintotarkkuutta sekä ajanjaksoa, jolta data on kerätty. Jokaisesta tutkimusaineistoon kuuluneesta avoimesta datalähteestä koottiin erillinen taulukko, johon kerättiin kootusti tieto kaikkien kriteerien toteutumisesta. Taulukot ja arvioiduista avoimen datan lähteistä löytyvät tiedot on kuvattu liitteestä 1. Taulukkoon lisättiin myös lähteet jokaiselle arvioidulle kriteerille, jonka avulla tutkimusprosessi pyrittiin pitämään mahdollisimman läpinäkyvänä.

Viimeisenä vaiheena arvioituja datalähteitä ja kriteerien täyttymistä analysoitiin tarkemmin vertailemalla datalähteitä toisiinsa ja määrittämällä mitkä avoimen datan hyödyntämisen haasteet nousivat useimmiten esille tutkimusaineistossa. Vertailun teki mahdolliseksi jokaisesta aineistosta koottu erillinen taulukko, johon oli kerätty samoilla kriteereillä kaikkien tutkimusaineistona olleiden avoimien datalähteiden kriteereiden täyttyminen. Taulukosta pystyttiin siten esimerkiksi helposti määrittämään, miten metadatat oli eri data-aineistoissa kuvattu tai oliko niitä määritelty ollenkaan. Vertailun perusteella pystyttiin määrittämään avoimen datan hyödynnettävyyden haasteet, jotka perusteltiin jokaisen mittaristoon kuuluvan arviointikriteerin perusteella tehtyjen havaintojen avulla.

Menetelmänä teoriaohjaavaa sisällönanalyysia hyödynnettiin erityisesti datan arvioinnissa, kun jokainen aineistoon kuuluva avoin datalähde arvioitiin edellä mainittujen kriteerien mukaisesti. Sisällönanalyysin avulla aineisto käytiin lävitse systemaattisesti, kun jokainen aineisto arvioitiin samojen kriteerien mukaisesti. Aineisto myös järjestettiin selkeään ja tiiviiseen muotoon, jotta datalähteiden välinen vertailu olisi mahdollisimman yksiselitteistä. Datalähteiden välistä vertailua tehtiin havainnoimalla millaisia haasteita kuinkin avoimen datalähteen hyödyntämiselle esiintyi ja sen jälkeen kartoittamalla mitkä haasteet esiintyivät aineistossa useasti. Datalähteiden vertailun tuloksena havaittiin aineistosta selkeitä tekijöitä, jotka aiheuttivat haasteita avoimen datan hyödyntämiselle globaalin organisaation big data analytiikassa.

5. TULOKSET

Avoimen datan hyödynnettävyyttä määrittävä mittaristo käsitti yhteensä kymmenen erillistä kriteeriä, joita arvioitiin laadullisin menetelmin. Mittaristossa korostettiin eri datalähteiden yhteentoimivuutta ja uudelleenkäytettävyyttä edistäviä tekijöitä, jotta avoimia data-aineistoja pystyttäisiin hyödyntämään mahdollisimman tehokkaasti globaalin organisaation big data analytiikassa. Hyödynnettävyyden haasteita on tutkittu kartoittamalla arviointikriteereiden toteutumista ja nostamalla esille tärkeimpiä tekijöitä.

Tutkimuksen tulokset, eli tutkimusaineistosta tehdyt havainnot avoimen datan hyödyntämisen haasteista globaalin organisaation big data analytiikassa, saatiin käymällä jokainen aineistoon kuuluva avoin datalähde läpi kaikkien mittaristoon kuuluvan arviointikriteereiden avulla. Tämän jälkeen keskeiset hyödyntämisen haasteet nostettiin esille kokoamalla tutkimuksen merkittävimmät havainnot ja niiden erityispiirteet yhteen taulukkoon. Tuloksissa korostuvat haasteet, jotka nähtiin esiintyvän useasti tai jotka aiheuttivat merkittävää haittaa avoimen data-aineiston hyödyntämiselle.

5.1 Saatavuus

Saatavuutta tarkasteltiin aineistosta sen perusteella, miten helposti aineisto oli löydettävissä ja käyttäjän hyödynnettävissä. Saatavuuden suurimpana esteenä nähtiin erillisten rekisteröintien ja kirjautumisten aiheuttama vaiva, joita avoimen datan tarjoajat vaativat datan loppukäyttäjiltä. Esimerkiksi Suomen ilmatieteenlaitos ja Euroopan karttatietoa tarjoava EuroGeographics vaativat käyttäjää kirjaamaan omat tietonsa ja hyväksymään lisensointisopimuksen ennen kuin aineistoon on mahdollista päästä lainkaan käsiksi. Puolestaan maailmanlaajuista ilmasto- ja säädataa tarjoava NCEI vaatii erillisen rekisteröinnin pelkästään ohjelmointirajapinnan käyttöön. Erillinen rekisteröinti tai kirjautuminen ennen datan käyttöä on usein vaadittu sen vuoksi, että avoimen datan tarjoaja haluaa kartoittaa millaisiin tarkoituksiin dataa halutaan hyödyntää.

Aineiston saatavuutta rajoitti myös jossain tapauksissa se, että aineiston löytyminen ja siihen käsiksi pääsy ei ollut verkkosivustolla näkyvällä paikalla tai muuten helposti löydettäessä. Tällöin käyttäjä joutui useampaan kertaan selaamaan eri näkymiä, ennen kuin löysi verkkosivulta kohdan, mistä aineiston pystyi joko lataamaan erillisenä tiedostona tai pääsi siihen muuten käsiksi. Esimerkiksi Suomen maanmittauslaitoksen sivustolta etsittäessä karttatietoja on saatavilla useasta paikasta ja sivustolla on esimerkiksi mahdollisuus tutustua maanmittauslaitoksen maastokarttoihin, ilmakuviin tai kiinteistörekistereihin. Avoimeen dataan pääsee käsiksi vasta ammattilaisille suunnattujen verkkopalveluiden alta.

Osaltaan avoimen data-aineiston saatavuutta rajoitti myös aineistojen heikko otsikointi ja luokittelu, jolloin käyttäjä ei välttämättä löydä oikeaa aineistoa useiden vaihtoehtojen joukosta. Esimerkiksi ilmakehästä ja ilmastosta saatavilla olevaa dataa avannut Merra tarjoaa useita eri data-aineistoja, jotka on kuvattu lyhenteiden ja koodien avulla. Ilmanpaineista tietoa tarjoava aineisto on nimetty lyhenteellä ”MAI3CPASM”, joka ei sellaisenaan uudelle käyttäjälle kerro vielä mitään. Lyhenteen merkitys on kuvattu erillisessä ohjeistuksessa, mutta epäselvä otsikointi haittaa huomattavasti aineiston nopeaa ja tehokasta hyödyntämistä.

5.2 Kokonaisuus ja laatu

Avoimen data-aineiston kokonaisuutta ja laatua arvioitiin sillä perusteella, miten avoimen data-aineiston laatua ja kokonaisuutta on kuvattu käyttäjälle. Laadun tarkastelemisessa on keskitytty myös määrittelemään, mihin aineiston laadun kuvaus pohjautuu sekä prosessiin datan keräämisen taustalla. Kokonaisuutta tarkasteltiin kartoittamalla, miten täydellistä ja oikeellista saatavilla oleva data oli. Tähän vaikutti esimerkiksi tiedon puuttuminen, NULL-arvojen ja tyhjien arvojen määrä sekä pääsy vain tiettyyn osaan kokonaisaineistosta.

Suurin osa aineistosta kuvasi ylätasolla datan laatua kuvaamalla sen keräysprosessia. Aineistosta esimerkiksi maailmanlaajuista jää- ja lumidataa tarjoava National Snow and Ice Data Center ja säätietoa tarjoava Weather Underground määrittelevät prosessin, jonka data käy läpi ennen julkaisemista. Yhdenmukaisella keräysprosessilla ja sen läpinäkyvyydellä pyritään varmistamaan datan tasalaatuisuus. Esimerkiksi Suomen ilmatieteenlaitos kuvaa hyvinkin tarkalla tasolla datan keräämistä ja laadunvarmistusta, joka sisältää sekä automaattisen laaduntarkistuksen että manuaalisen määrityksen. Aineiston kuvaaminen on lähes välttämättä subjektiivista, jonka vuoksi yhtenevien laatukriteerien tai standardien puute aiheuttaa haasteen aineistojen laadun määrittämiselle. Suomen ilmatieteenlaitoksen ja Euroopan karttatietoa tarjoavan EuroGeographicsin aineistoissa datan laatua oli kuvattu erillisellä ISO-standardilla, joka osaltaan luo uskottavuutta aineiston laadun varmistamiseksi.

Osa aineistosta ei kuitenkaan ottanut mitenkään kantaa aineiston laadukkuuteen. Esimerkiksi avointa ja maailmanlaajuista karttatietoa tarjoava Natural Earth ei kuvaa mitenkään aineiston laadukkuutta tai datan keräysprosessia. Myöskään avointa säätietoa tarjoava OpenWeatherMap ei määrittele aineiston laadukkuutta itse millään tavoin, vaan luottaa datan alkuperäisten lähteiden määritelmiin datan laadusta. Palveluun on kuitenkin mahdollista liittää myös yksityisiä sääasemia ja niiden tuottamaa dataa, joiden tarjoaman datan laatua ei ole kuvattu millään tavalla.

Data-aineiston kokonaisuuteen ei kovinkaan monessa lähteessä otettu kantaa. Aineistojen kuvaus ei siten lähes poikkeuksetta sisällä tietoa siitä, onko kaikki mahdollinen data avattu vai pelkästään pieni osuus. Datalähteen kokonaisuutta pyrittiin tarkastelemaan

myös puutteellisten data-aineistojen määrällä, mikäli dataa oli mahdollista tarkastella tyhjien ja NULL-arvojen osalta. Esimerkiksi Euroopan ilmastotietoa tarjoaja European Climate Assessment & Dataset sisälsi hyvinkin puutteellisia aineistoja, sillä NULL-arvoja esiintyi lähes jokaisessa saatavilla olevassa aineistossa. Samaa haastetta ei kuitenkaan havainnoitu yhtä merkittävästi muiden aineistojen kohdalla, vaan tyhjiä arvoja oli vain yksittäisiä. Vain pieni osa avoimista datalähteistä määrittelee, että avoimena datana on maksutta saatavilla vain rajoitettu osa koko aineistosta. Esimerkiksi Open Weather Map tarjoaa maksua vastaan kattavampia tietoaineistoja, jolloin data-aineiston kokonaisuus voidaan kyseenalaistaa.

5.3 Dataformaatti

Dataformaatin osalta avoimia data-aineistoja tarkasteltiin kartoittamalla, missä eri muodoissa data-aineistoja oli mahdollisuus hyödyntää ja uudelleen käyttää. Aineistosta havaittiin, että avointa dataa on saatavilla huomattavan monessa eri formaatissa. Esimerkiksi maailman laajuista ympäristötietoa ja säädataa tarjoava NCEI tarjoaa dataa ASCII, ArcGIS, KMZ, PDF, CSV formaateissa, kun taas Suomen iltatieteenlaitokselta avointa säädataa on saatavilla vain XML-formaatissa. Puolestaan Yhdysvaltain ilmailu- ja avaruushallinto NASA tarjoaa ilmastosta ja säästä saatavaa avointa dataa HDF ja NetCDF formaateissa. Yhteensä aineistossa esiintyi useita kymmeniä eri dataformaatteja, joista esimerkiksi Pelkästään Yhdysvaltojen avoimen datan katalogi Data.gov tarjoaa ilmastoon liittyvää dataa yhteensä 48 eri formaatissa.

Avoimissa data-aineistoissa esiintyi myös jonkin verran kaupallisten tarjoajien formaatteja, kuten XLS, joka ei ole kaikkien yhtälailla hyödynnettävissä vaan vaatii maksullisen ohjelman ostamisen. Esimerkiksi Yhdysvaltojen energiaministeriö tarjoaa dataa vain XLS ja PDF-muodoissa. Muita XLS-dataformaattia käyttäviä avoimia datalähteitä ovat esimerkiksi Open Energy Information, Entso-E ja National Snow and Ice Data Center. Myös avointa paikkatietoa oli myös saatavilla useissa eri formaateissa, joista yleisimpiä olivat ESRI Shapefile, GeoTIFF sekä LAZ.

Yleisimpiä dataformaatteja koko aineistolle olivat ASCII, CSV ja XML, joista ASCII ja CSV ovat hyvin yksinkertaisia tekstitiedostaja kun taas XML on rakenteellinen kuvauskieli, joka auttaa jäsentämään laajoja data-aineistoja. Kaikki käytetyimmät dataformaatit ovat avoimia ja siten hyödynnettävissä riippumatta tietystä ohjelmistosta. Havainnoista käykin ilmi, että suurin haaste avoimien data-aineistojen hyödynnettävyydessä on hyvin suuri määrä eri dataformaatteja eikä niinkään formaattien avoimuuden puute.

5.4 Arkkitehtuuri ja rajapinnat

Avoimien data-aineistojen arkkitehtuurikuvauksia ja rajapintoja tarkasteltiin määrittämällä miten datan alkuperäiset tietovarannot ja niiden arkkitehtuuri on kuvattu, sekä millainen rajapinta datan hyödyntämiselle on tarjottu. Arkkitehtuurikuvausten määritte-

lyssä korostettiin erityisesti aineistojen alkuperäisten datalähteiden kuvaamista, sillä useat avointa dataa tarjoavat lähteet ovat koonneet dataa useammasta eri lähteestä. Rajapintoja kartoitettiin määrittämällä erityisesti ohjelmointirajapintojen hyödyntämisen mahdollisuutta.

Avoimen datan arkkitehtuuria ei oltu suurimmassa osassa datalähteitä kuvattu juuri ollenkaan. Esimerkiksi Suomen ilmatieteenlaitos, karttatietoa tarjoavat Eurogeographics ja Natural Earth, Euroopan sähköverkkotietoa tarjoava Entso-E, Yhdysvaltojen energiaministeriö sekä Suomen maanmittauslaitos eivät kuvaa millään tavalla avaamansa datan arkkitehtuuria. Mikäli datalähteen arkkitehtuuria on aineistossa kuvattu, rajoittuu määrittely data-arkkitehtuurin kuvaukseen alkuperäisistä datalähteistä. Esimerkiksi globaalia säätietoa tarjoava Open Weather Map kuvaa arkkitehtuuriaan määrittämällä listauksen datojen alkuperästä sekä maininnan, että dataa kerätään eri maiden ilmatieteenlaitosten lisäksi yli 40 000 yksityiseltä sääasemalta. Samoin määrittelee avointa ilmastodataa tarjoava Merra, joka listaa 12 eri datalähdettä, jotka tuottavat tarjottavan datan.

Data-aineistojen rajapintoja oli suurimmassa osassa aineistoa kuvattu ainakin jollain tasolla. Yleisimpiä rajapintoja avoimen datan hyödyntämiselle olivat ohjelmointirajapinnat sekä FTP-protokollan hyödyntäminen, joista esimerkiksi Euroopan ilmastotietoa tarjoava European Climate Assessment & Dataset käyttää rajapintana perinteistä FTP-tiedonsiirtoa. Osa avoimista datalähteistä tarjosi myös helpomman tavan datan käsittelyyn tarjoamalla erillisen ohjelmointirajapinnan. Esimerkiksi maailmanlaajuista ilmastoja säädataa tarjoava NCEI ja globaalia säädataa tarjoavat OpenWeateherMap sekä Weather Underground tarjoavat ohjelmointirajapinnan, jolloin aineistoa voidaan käyttää suoraa rajapinnan yli reaaliaikaisesti JSON tai XML -muodossa. Myös wiki-tekniikalla toimiva Open Energy Information tarjoaa erillisen OpenEI REST-rajapinnan. Suurin osa aineistosta ei kuitenkaan tarjoa helppokäyttöistä ohjelmointirajapintaa.

5.5 Käyttöehdot

Avoimien datalähteiden käyttöehtoja tarkasteltiin kartoittamalla, miten aineiston käyttöehtoja ja uudelleenkäyttöä on kuvattu. Aineiston käyttöehtojen kuvaamisessa painotettiin erityisesti uudelleenkäytön ja hyödynnettävyyden mahdollisuutta. Tällöin käyttäjälle tulee olla esitettynä selkeästi ja läpinäkyvästi käyttöehdot, jotka kannustavat datan uudelleenkäyttöön.

Aineistosta havaittiin, että suurin osa avoimesta datasta on vapaasti kaikkien hyödynnettävissä eikä käyttöä ole pyritty rajoittamaan. Ainoastaan European Climate Assessment & Dataset sekä Weather Underground aineistojen käyttö oli rajoitettu pelkästään eikaupallisen tutkimuksen ja opetuksen käyttöön. Haasteena kuitenkin nähtiin, ettei suurimmassa osassa aineistoa käyttöehtoja oltu kovinkaan selkeästi esitetty käyttäjälle. Ainoastaan avointa paikkatietoa tarjoava Natural Earth tuo käyttäjälle selkeästi tietoon, että dataa saa hyödyntää millä tavalla tahansa, mukaan lukien sisällön muokkaamisen ja

uudelleen levittämisen opetukseen, henkilökohtaisiin tai kaupallisiin tarkoituksiin. Myös Suomen maanmittauslaitos korostaa käyttöehdoissa, että avoimet data-aineistot ovat vapaasti kaikkien hyödynnettävissä laajan ja pysyvän käyttöoikeuden myötä.

Käyttöehtojen lisäksi jotkin avoimet datalähteet vaativat erilliset lisenssin aineiston hyödyntämiseen. Esimerkiksi Suomen maanmittauslaitos, kuten myös ilmatieteenlaitos ja Euroopan karttatietoa tarjoava EuroGeographics vaativat erillisen lisenssin hyväksynnän. Maanmittauslaitos ja ilmatieteenlaitos vaativat CC 4.0 -lisenssiä, joka on kansainvälinen avoimen datan käyttöehtojen määrittelemiseen käytetty lisenssi. Se sallii datan jakamisen ja muokkaamisen käyttäjän omia käyttötarkoituksia varten. Ainoastaan lähde alkuperäiseen dataan on mainittava. EuroGeographics puolestaan hyödyntää omaa avoimen datan lisenssiään, joka sisältää pääpiirteittäin kuitenkin täysin samat asiat, kuin kansainvälinen CC 4.0 -lisenssi.

Maininta avoimen datan alkuperäisestä lähteestä nähtiin useammassakin aineistossa käyttöehdon takeena. Näin määrittelee käyttöehdoissaan esimerkiksi globaalia lumi- ja jäätietoa tarjoava National Snow and Ice Data Center, ilmastotietoa tarjoava Merra sekä Yhdysvaltain avoimendatan katalogi Data.gov ja Euroopan sähköverkkotietoa tarjoava Entso-E.

5.6 Kustannukset

Avoimien datalähteiden hyödyntämisen kustannuksia kartoitettiin määrittämällä, onko aineistojen käyttö ilmaista vai peritäänkö sen käytöstä maksua. Aineistosta pyrittiin myös tarkastelemaan, millä tavalla käytönkustannukset on käyttäjälle ilmaistu vai onko kustannuksia kuvattu millään tavalla. Maksullisten aineistojen kohdalla pyrittiin myös kartoittamaan, onko hinnasto kuvattu selkeästi käyttäjälle ja edelleen selvittämään mistä kustannukset aineiston käytölle muodostuvat sekä onko hinnoittelu tehty kustannus- vai katepohjaisesti.

Avoimien datalähteiden hyödyntäminen oli suurimmassa osassa aineistoja täysin ilmaista, eikä käytöstä peritty mitään maksuja. Suurimmassa osassa avoimen aineiston maksuton hyödyntäminen oli esitetty selkeästi muiden käyttöehtojen mukana ja esimerkiksi maailmanlaajuista karttatietoa tarjoava Natural Earth sekä Suomen säätietoa tarjoava ilmatieteenlaitos kertovat selkeästi heti etusivulla, että aineistot ovat saatavilla maksutta julkiseen käyttöön. Käyttäjälle tieto aineiston käytöstä muodostuvista kustannuksista on ensisijaisen tärkeää, jonka vuoksi maksuttomuutta olisi tärkeää korostaa selkeästi kaikilla sivustoilla.

Kuitenkin muutama avointa dataa tarjoava sivusto tarjosi kattavampia aineistoja maksua vastaan. Esimerkiksi globaalia säätietoa tarjoava Weather Underground tarjoaa ohjelmointirajapinnan vain maksua vastaan ja samoin OpenWeatherMap tarjoaa vain nykyisen säätiedon sekä kolmen päivän ennusteen ilmaiseksi, mutta pidemmän aikavälin en-

nusteen tai historiallisen säädatan saa käyttöönsä vain erillisellä kuukausimaksulla. Avoimen datan tarjoaminen nähtiin siten edellä mainituissa palveluissa enemminkin keinona houkutella myös maksavia käyttäjiä ja kustannukset perustuvat puhtaasti katehinnoitteluun, kun datan avulla halutaan harjoittaa kannattavaa liiketoimintaa.

5.7 Metadatan kuvaus

Aineistojen metadatan osalta avoimia datalähteitä tarkasteltiin kartoittamalla miten selkeästi niiden metadatat on kuvattu. Metadatan kuvaamisessa tarkasteltiin miten tarkalla tasolla metadataa on saatavilla ja millaisia tietoja siinä on kuvattu. Kaikkien aineistojen kohdalla myös määriteltiin, onko kuvausta aineiston metadatasta ylipäätään tehty ollenkaan.

Avoimien data-aineistojen metadatan kuvaamista kartoittaessa havaittiin, että metadatan kuvaamisessa sekä sisällössä on huomattavia eroja eri aineistojen sekä datan tarjoajien kesken. Esimerkiksi säädataa tarjoavat OpenWeatherMap sekä Weather Underground sekä sähköverkkodataa tarjoavat Entso-E ja Yhdysvaltain energiaministeriö eivät kuvaa aineistojensa metadataa millään tavalla. Puolestaan Suomen ilmatieteenlaitos, Yhdysvaltain avoimen datan katalogi Data.gov, Suomen maanmittauslaitos, karttatietoa tarjoava EuroGeographics sekä GeoPlatform ja Global Atlas for Renewable Energy määrittävät hyvin tarkalla tasolla kaikkien aineistojensa kuvaukset ja metadatan. Näistä datan tarjoajista Yhdysvaltain Data.gov sekä GeoPlatform noudattavat ISO-19139 -standardin mukaista metadatan kuvausta paikkatiedolle.

Suuri osa aineistosta kuvasi metadatan tietoja esimerkiksi määrittämällä sanallisen kuvauksen aineiston sisällöstä, versiohistoriasta, maantieteellisestä tai ajallisesta kattavuudesta. Esimerkiksi maailmanlaajuista ilmastodataa tarjoava NCEI tarjoaa lyhyen kuvauksen jokaisen aineiston sisällöstä ja puolestaan Euroopan ilmastotietoa tarjoava Europen Climate Assessment & Dataset ei tarjoa erikseen minkäänlaista metadatan kuvausta, vaan ainoastaan ladattavan aineiston alkuun on merkitty käytettävien lyhenteiden selitteet. Myös maailmanlaajuista ilmastodataa tarjoava Merra kuvaa jokaisen aineiston sisällön ja käytetyt merkinnät lyhyesti, mutta erillistä metadataa ei ole saatavilla. Myös oleellisia tietoja, kuten kuvaus aineiston laadusta, keräysprosessista tai käytetyistä mittaustekniikoista.

5.8 Maantieteellinen alue

Avoimien datalähteiden maantieteellistä kattavuutta tarkasteltiin määrittämällä millaisen alueen data kattaa, eli onko tietoa saatavilla globaalisti, maanosakohtaisesti, valtiokohtaisesti tai edelleen rajatummalla maantieteellisellä alueella. Maantieteellisen alueen kattavuutta tarkasteltiin erityisesti kartoittamalla miten laajoja aineistoja on saatavilla ja joudutaanko dataa yhdistelemään useammasta eri lähteestä globaalin organisaation tarpeisiin.

Datan kattama maantieteellinen alue oli määritelty lähes jokaisessa aineistossa selkeästi. Havaintona huomattiin, että avointa dataa on saatavilla useimmin vain rajatulla maantieteellisellä alueella, kuten yleisimmin valtiokohtaisesti. Luonnollisesti Suomen ilmatieteenlaitoksen sekä maanmittauslaitoksen aineistot kattavat vain Suomen maantieteellisen alueen ja samoin Yhdysvaltojen avoimen datan katalogi Data.gov kattaa tarkastelussaan tarkimmin vain Yhdysvaltojen maantieteellisen alueen. Tämän lisäksi aineistoista European Climate Assessment & Dataset, EuroGeographics sekä Entso-E kattavat tarkastelussaan hieman laajemman maantieteellisen otannan, kun ne kattavat koko Euroopan.

Aineistosta globaalin tarkastelun mahdollistivat avointa säädataa tarjoavat ilmastotietoa tarjoava NCEI, lumi- ja jäädataa tarjoava National Snow and Ice Data Center, OpenWeatherMap sekä Weather Underground ja Merra. Paikkatietoa tarjosi vain Natural Earth sekä tuuliturbiinitietoa tarjoava Global Atlas for Renewable Energy. Tarkastelussa kuitenkin havainnoitiin, että suurin osa globaalista avoimesta datasta kattoi kuitenkin vain keskiarvoja tai pieniä otoksia useasta sijainnista eikä laajaa näkemystä globaalille tarkastelulle ollut yksinkertaista toteuttaa. Tarkemman tason tarkastelu koko globaalille maantieteelliselle alueelle on havaintojen mukaan lähes välttämätön toteuttaa yhdistämällä useita eri aineistoja ja linkittämällä niiden data toisiinsa.

5.9 Havaintotarkkuus

Avoimien data-aineistojen havaintotarkkuuden kartoittamisessa tarkasteltiin millaisella havaintotarkkuudella erityisesti säädataa oli saatavilla. Havaintotarkkuuden tarkastelussa kiinnitettiin myös erityisesti huomiota, oliko saatavilla tarkkoja data-arvoja ja millä tiheydellä ne on mitattu, vai onko esimerkiksi tiedoista saatavilla vain keskiarvoja tarkkojen arvojen sijaan. Havaintotarkkuuden kuvaamisessa kartoitettiin myös, miten data-arvot oli määritelty ja onko havaintotarkkuuksien osalta nähtävissä mitään yhteistä tekijää, joka mahdollistaa eri aineistojen vertailun keskenään.

Säädataa tarjoavista avoimista datalähteistä suurimmassa osassa oli selkeästi määritelty datan havaintotarkkuus, eli miten usein arvot on kerätty. Esimerkiksi globaalia ilmastotietoa tarjoava NCEI kerää dataa tunnin, päivän, kuukauden, vuoden ja useamman vuoden tarkkuuksilla. Lähes samaa havaintotarkkuutta käyttää avointa ilmastodataa tarjoava Merra, jossa dataa on saatavilla päivän ja kuukauden tarkkuuksilla. Tarkimmalla havaintotarkkuudella dataa on avannut Suomen ilmatieteenlaitos, jossa uusinta dataa on saatavilla 10 minuutin tarkkuudella ja vanhempiakin aineistoja vähintään päivittäisellä tarkkuudella.

Huomioitavaa havaintotarkkuuksia tarkasteltaessa onkin, että suurin osa avoimista data-aineistoista tarjoaa lähes ainoastaan keskiarvoja datasta, eikä tarkkoja mittausarvoja. Esimerkiksi Euroopan ilmastodataa tarjoava European Climate Assessment & Dataset on avannut dataa 75 sääasemalta määrittäen muutokset keskiarvoissa ja ääriarvot ilmas-

tossa. Myös globaalia säädataa tarjoava Weather Undergound määrittää ainoastaan säätiedot päivittäisellä tasolla kattamalla arvon päivän keskilämpötilasta sekä alimmasta ja korkeimmasta mitatusta arvosta. Vain osalle aineistosta on mahdollista saada dataa tunnin tarkkuustasolla. Aineistosta havaittiin, että säätietojen vertailu eri aineistojen kesken on havaintotarkkuuksista johtuvien eroavaisuuksien vuoksi haastavaa, sillä dataa on saatavilla hyvin erilaisilla tarkkuustasoilla. Paikkatietoa tarjoavien datalähteiden osalta havaintotarkkuutta ei oltu juurikaan määritelty, sillä paikkatiedot eivät muutu läheskään yhtä useasti kuin esimerkiksi säätiedot.

5.10 Ajanjakso

Ajanjaksoa tarkasteltaessa avoimista data-aineistoista kartoitettiin, miltä aikaväliltä dataa oli saatavilla. Ajanjaksoja tarkasteltaessa kiinnitettiin erityistä huomiota siihen, miten usealta vuodelta historiaan perustuvaa dataa oli saatavilla ja miten aineiston kattama ajanjakso oli aineistossa kuvattu. Tämän lisäksi ajanjakson tarkastelussa kartoitettiin datan ajantasaisuutta sekä päivitystiheyttä.

Data-aineistojen ajanjaksot oli pääosin kuvattu aineistoissa selkeästi. Suurin osa säädataa tarjoavista aineistoista kattoi historiallista dataa useamman vuosikymmenen takaa, jopa 1800-luvulta lähtien. Esimerkiksi maailmanlaajuista ilmastodataa tarjoava NCEI, Yhdysvaltain ilmastotietoa tarjoava Data.gov sekä lumi- ja jäädataa tarjoava National Snow and Ice Data Center ovat avanneet dataa 1800-luvulta tähän päivään asti. Puolestaan Euroopan ilmastodataa tarjoava European Climate Assessment & Dataset on avannut dataa vuodesta 1918 alkaen ja globaalia säädataa tarjoava Weather Underground vuodesta 1945 alkaen. Ainoastaan sähköverkkodataa tarjoava Entso-E on avannut dataa vasta vuodesta 2009 asti ja kuten myös Yhdysvaltojen energiaministeriö, joka tarjoaa dataa vuodesta 2000-alkaen. Säädataa tarjoavat OpenWeatherMap sekä Weather Underground tarjoavat historiallisen datan lisäksi dataa sääennusteista.

Avoimen datan ajantasaisuutta kuvattiin ei aineistoissa hyvin eri tasolla ja vain muutama avoimen datan tarjoaja ilmoittaa selkeästi milloin aineisto on julkaistu sekä miten usein sitä päivitetään. Esimerkiksi Suomen maanmittauslaitos päivittää kaikki aineistonsa vähintään vuosittain, mutta joitain aineistoja päivitetään jopa viikkotasolla. Myös Yhdysvaltojen paikkatietoa tarjoava GeoPlatform päivitetään vähintään kuukausittain. Kuitenkaan läheskään kaikki aineistojen tarjoajat eivät kuvanneet miten usein aineistoja päivitetään tai miten usein uutta dataa on saatavilla.

5.11 Tunnistetut avoimen datan hyödynnettävyyden haasteet

Mittaristosta havaittiin arviointikriteerien avulla selkeitä avoimen datan hyödynnettävyyden haasteita globaalin organisaation big data analytiikalle. Suuri osa tunnistetuista haasteista esiintyy myös yleisemmin avoimen datan hyödynnettävyyttä pohdittaessa, mutta korostui erityisesti tämän tutkimuksen kontekstissa. Mittaristosta nousi esille seit-

semän erillistä hyödynnettävyyden haastetta, jotka koettiin esiintyvän useimmin tai niiden koettiin aiheuttavan suurimman haasteen avoimen datalähteen uudelleenhyödyntämisen, aineistojen yhdistelemisen ja big data analytiikan näkökulmista. Haasteiksi tunnistettiin tiedostomuotojen eroavaisuudet, puutteet metadatassa, eroavaisuudet havaintotarkkuuksissa, maantieteelliset rajoitteet, heikko arkkitehtuurikuvaus ja rajapinnat, eroavaisuudet datan laadussa sekä heikko saatavuus ja löydettävyys. Toisaalta tutkimuksen tuloksena havaittiin myös, että arviointikriteereistä kustannuksia ja ajanjaksoa ei nähty merkittävänä haasteena avoimen data-aineiston uudelleenhyödyntämiselle. Pääosin avointen datalähteiden ajanjaksot oli kuvattu selkeästi ja lähes kaikki avoin data oli saatavilla ilman kustannuksia. Taulukkoon 5 on koottu tutkimusaineistosta tunnistetut avoimen datan hyödynnettävyyden haasteet sekä erityispiirteet niiden esiintymiselle.

Taulukko 5. *Avoimen datan hyödynnettävyyden haasteet*

Tiedostomuotojen eroavaisuudet	• Dataa saatavilla useissa eri formaateissa • Tarve useiden ohjelmistojen käyttöön • Aineistojen yhdisteleminen vaatii usein manuaalista työtä
Puutteet metadatassa	• Termien ja lyhenteiden käyttö • Mahdollisuus väärinymmärrykselle • Tulkinnanvaraisuus
Erot havainto-tarkkuuksissa	• Havaintoja useilla eri tarkkuustasoilla • Yhdisteltävyys muihin data-aineistoihin haastavaa
Maantieteelliset rajoitteet	• Data kattaa usein vain rajatun maantieteellisen alueen • Tarkastelualueen ollessa laajempi, ei dataa usein ole saatavilla tarvittavalla tarkkuudella
Heikko arkkitehtuurikuvaus ja rajapinnat	• Ei aina tietoa kuka on kerännyt datan tai mistä se on peräisin • Puutteita rajapinnoissa, vain osa tarjoaa hyvän ohjelmointirajapinnan
Eroavaisuudet datan laadussa	• Puutteellinen laadun kuvaaminen • Standardien puute
Heikko saatavuus ja löydettävyys	• Erillisten rekisteröintien vaiva • Aineistojen huono löydettävyys ja puutteet nimeämisissä

Tutkimuksessa havaittiin, että avointa dataa on saatavilla huomattavan monessa eri dataformaatissa ja tiedostomuodot saattoivat vaihdella jopa saman datan tarjoajan kohdalla useiden kymmenien eri formaattien välillä. Kartoittamalla aineistosta löytyviä eri tiedostomuotoja huomattiin, että avointa dataa on saatavilla esimerkiksi CSV ja ASCII muodoissa, jotka täyttävät dataformaatteina esimerkiksi Berners-Leen (2006) asettamat vaatimukset avoimelle datalle. Samanaikaisesti dataa oli kuitenkin saatavilla myös PDF ja HTML -muodoissa, jotka puolestaan ei millään muotoa täytä avoimelle datalle asetettuja vaatimuksia. Kun aineistoja on saatavilla hyvin useissa eri muodoissa, tuottaa se haasteita myös eri aineistojen yhdistelemiselle ja linkittämiselle keskenään. Myös dataformaatit, jotka eivät ole koneluettavassa muodossa, aiheuttavat huomattavan määrän työtä jotta aineistoja voitaisiin tehokkaasti hyödyntää uudelleen. Havaintojen perusteella voidaankin todeta, että data-aineistojen yhdisteleminen vaatii hyvin suuren määrä manuaalista työtä, kun data-aineistojen formaatit halutaan yhtenäistää. Tämä voidaan usein toteuttaa erillisten muuntimien tai ohjelmistojen avulla, mutta tällöin on huomioitava myös virheiden mahdollisuus. Mitä enemmän aineistoa joudutaan muokkaamaan ja käsittelemään ennen hyödyntämistä, sitä riskialttiimpaa on datan vääristyminen, hukkuminen tai muuten virheellisten arvojen syntyminen. Vaihtelevat tiedostomuodot vaativat myös usein monien eri ohjelmistojen käyttöä. Esimerkiksi shapefile-tiedostona paikkatietoa tarkasteltaessa tarvitaan erillinen paikkatietojärjestelmä kuten Esri, jolla dataa voidaan edelleen hyödyntää ja tarkastella, kun taas puolestaan XLS-muodossa tarjottavaa dataa pystyy hyödyntää vasta ostamalla kaupallisen Microsoft Excel -ohjelmiston. Avoimelle datalle ei aineiston perusteella ole yhteisesti määritetty yhtä tai useampaa tiedostomuotoa, joita datan tuottajien suositeltaisiin hyödyntävän, vaan datan tarjoajat määrittävät täysin itse, millaisessa dataformaatissa haluavat datan avata. Kaikkea dataa ei ymmärrettävästi voida tarjota yhtenevässä tiedostomuodossa, mutta erilaisten tiedostomuotojen vähentäminen ja avoimen datan formaattien yhdenmukaistaminen helpottaisi merkittävästi avoimen datan hyödyntämistä osana big data analytiikkaa.

Toisena merkittävänä avoimen datan hyödyntämisen haasteena tunnistettiin aineistojen metadatan puutteellinen kuvaaminen. Vain osassa aineistoja metadataa oli kuvattu kattavasti, usein näissäkin tapauksissa kuvauksissa esiintyneet tiedot poikkesivat keskenään huomattavasti eikä vertailua eri aineistojen välillä voitu suorittaa metadatan tarjoamien tietojen perusteella. Tutkimusaineistosta ei siten havaittu yhteistä kuvausta avoimien data-aineistojen metadatalle, vaan kuvaus sisälsi usein vain lyhyen sanallisen määritelmän data-aineiston keskeisestä sisällöstä. Esimerkiksi käytettyjen lyhenteiden ja merkintöjen selitteiden puuttuminen haittaa merkittävästi aineiston uudelleenhyödyntämisen mahdollisuuksia, sillä käyttäjä ei välttämättä ymmärrä datan kuvaamia tekijöitä. Väärinymmärrykset ja tulkinnanvaraisuus data-aineistosta edelleen kasvattavat virheellisten analyysien mahdollisuutta sekä väärien johtopäätösten tekemistä. Huomattavaa oli myös, että osa tutkimusaineistona olleista avoimista datalähteistä ei tarjonnut minkään tasoista metadatan kuvausta aineistojen loppukäyttäjille. Metadatan kuvaaminen voidaan nähdä välttämättömänä avoimen data-aineiston uudelleenhyödyntämisen kannalta,

sillä ymmärtämättömyys aineiston sisällöstä asettaa useissa tapauksissa lähes ylitsepääsemättömän esteen avoimen datan hyödyntämiselle kun aineistosta ei voida tuottaa analyysejä tai niiden tulokset ymmärretään väärin.

Aineistojen yhdisteltävyyttä tarkasteltaessa huomattiin myös, että aineistojen havaintotarkkuudet poikkesivat keskenään huomattavan paljon ja dataa oli saatavilla useilla eri tarkkuustasoilla. Esimerkiksi säädataa oli saatavilla jossain aineistossa tarkkoina mittausarvoina 10 minuutin välein, kun taas toinen aineisto tarjosi pelkästään päivän tai kuukauden keskiarvon. Tällöin aineistojen välistä vertailua tai yhdistelemistä ei voida tehdä sellaisenaan, vaan tarkemman tarkastelun mahdollistavaa aineistoa on väistämättä muokattava vähemmän tarkalle tasolle. Havaintotarkkuuksia kartoitettaessa havaittiinkin, että aineistojen yhdistäminen vaati lähes poikkeuksetta aineistojen tarkkuustason laskua, sillä data-arvot eivät sellaisenaan useinkaan olleet vertailukelpoisia keskenään vaan vaativat muokkaamista yhteneväiselle havaintotarkkuudelle. Erot havaintotarkkuuksissa ja aineistojen tarkkuustasoissa muodostavat haasteen eri aineistojen yhdistelemiselle keskenään eikä tarkkoja analyysejä voida välttämättä toteuttaa ollenkaan. Datan tarkkuustason lasku ei useinkaan merkittävästi vaikuta pidemmän aikavälin trendien tai muun kehityksen seurantaan, mutta havaintotarkkuudella on hyvin suuri merkitys mikäli halutaan analysoida dataa päivittäisellä tai yhä tarkemmalla tarkkuustasolla.

Avoimien data-aineistojen havaittiin myös olevan lähes poikkeuksetta maantieteellisesti rajoittuneita, vaikka tutkimusaineistossa oli erityisesti painotettu sellaisia datalähteitä, jotka tarjoavat dataa mahdollisimman laajalla maantieteellisellä otannalla. Suurin osa avoimista datalähteistä olikin julkisten organisaatioiden tarjoamia, jonka vuoksi tarkastelu usein rajoittui joko valtiotasolle tai yhteen maanosaan. Vain osa aineistosta tarjosi avointa dataa, joka kattaa globaalisti kaikki maantieteelliset alueet. Tutkimuksen avulla havaittiin myös, että mikäli aineisto tarjosi avointa dataa globaalilla maantieteellisellä kattavuudella, ei dataa useinkaan oltu kuvattu tarvittavalla tarkkuustasolla. Esimerkiksi historiallisista säätiedoista oli havaintoja usein vain muutamista sijainneista ja käyttäjälle tarjottiin vain keskiarvoja tai havaittuja maksimi- tai minimiarvoja. Maantieteellisesti rajoittuneet aineistot vaativat siten lähes poikkeuksetta useiden eri aineistojen yhdistämistä, mikäli dataa halutaan tarkastella globaalisti tarvittavalla tarkkuustasolla. Aineiston maantieteellistä rajoittuneisuutta ei välttämättä nähtäisi haasteena, mikäli eri maat tarjoaisivat yhtenevät avoimen datan aineistot keskenään, jolloin myös vertailu on mahdollista. Havaintojen perusteella maantieteellinen rajoittuneisuus asettaa merkittävän haasteen eri aineistojen yhdistämiselle ja analytiikan mahdollisuuksille, kun käyttäjä haluaa tarkastella arvoja mahdollisimman tarkalla tasolla.

Tarkasteltaessa tutkimusaineostoon kuuluvien avoimien data-aineistojen kuvausta arkkitehtuurista ja tarjotuista rajapinnoista, havaittiin selkeitä puutteita arkkitehtuurien kuvaamisessa sekä rajapinnoissa, jotka merkittävästi heikentävät aineiston uudelleenkäytettävyyttä. Heikosta arkkitehtuurin kuvaamisesta johtuen käyttäjä ei välttämättä tiedä miten ja mistä alkuperäisistä datalähteistä saatavilla oleva avoin data on peräisin.

Puutteellinen arkkitehtuurin kuvaus aiheuttaa siten haasteita aineiston uudelleenhyödyntämisen kannalta, kun aineiston tietorakennetta ei osata tulkita oikein. Hyödynnettävyyden haasteeksi tunnistettiin myös ohjelmointirajapintojen puuttuminen, sillä vain osa aineistosta tarjosi suoran ohjelmointirajapinnan aineiston hyödyntämiseksi. Ohjelmointirajapinnan puuttuminen lisää merkittävästi manuaalista työtä, kun aineistoja joudutaan siirtelemään paikasta toiseen aina kun aineistoon tulee päivityksiä tai lisäyksiä. Samalla reaaliaikaisen big data analytiikan toteuttaminen hankaloituu merkittävästi, kun uutta aineistoa ei pystytä hyödyntämään heti sen ilmestyttyä. Tällöin myös jatkuvan ja pitkäaikaisen data-analytiikan hyödyntäminen hankaloituu, kun analyysejä ei voida automatisoida, vaan uusi aineisto joudutaan liittämään manuaalisesti osana erillistä työtä.

Aineistosta havaittiin myös selkeitä eroavaisuuksia avoimen datan laadussa ja erityisesti laadun kuvaamisessa. Puutteellinen data-aineiston laadun kuvaaminen vaikuttaa kielteisesti aineiston läpinäkyvyyteen ja luotettavuuteen, kun käyttäjä ei pysty varmistumaan saatavilla olevan datan paikkansapitävyydestä. Laadun kuvaaminen on lähes poikkeuksetta subjektiivista, jonka vuoksi keskenään eroavat kuvaukset voivat aiheuttaa merkittäviä haasteita tai jopa esteitä aineiston uudelleenhyödyntämiselle. Eroavaisuudet aineiston laadun kuvaamisessa aiheuttavat helposti tilanteita, joissa laadukkuutta ei pystytä selkeästi määrittämään eikä tämän vuoksi saatavilla olevaa avointa data-aineistoa voida välttämättä yhdistellä muihin aineistoihin tai hyödyntää lainkaan. Aineistojen laadun kuvaamiselle ei ole olemassa tiettyä standardia tai määritelmää, joka kattaisi tarvittavat kuvaukset data-aineiston laadusta. Standardin puute aiheuttaa myös ongelman datan hyödyntäjän näkökulmasta, mikäli kaikkia tarpeellisia laatuun liittyviä osatekijöitä ei ole kuvattu. Useat aineistot eivät kuvanneet aineiston laatua ollenkaan ja monien aineistojen laadunkuvaamisesta puuttui myös laadunvarmentamiseen sekä datankeräämiseen liittyvät määritelmät. Kun aineiston laadukkuuteen ei oteta kantaa mitenkään, on käyttäjän hankala varmistua siihen, että tämä pystyy luottamaan aineistosta saataviin analyyseihin ja niiden tuloksiin.

Merkittävänä avoimen datan hyödyntämisen haasteena havaittiin myös avoimen datan heikko saatavuus sekä löydettävyys. Data-aineistojen saatavuuden haasteena koettiin erityisesti erillisten rekisteröintien vaiva. Käyttäjiltä vaadittiin usein erillinen rekisteröinti tai kirjautuminen, jonka avulla dataa tarjoava organisaatio voi varmistua käyttöehtojen täyttymisestä tai lisenssiehtojen hyväksymisestä sekä pyrkii samalla saamaan tietoa mihin käyttötarkoitukseen avointa dataa halutaan hyödyntää. Kuitenkin dataa hyödyntävän käyttäjän näkökulmasta erillinen rekisteröinti tai kirjautuminen aiheuttaa tarpeetonta vaivaa, kun data-aineistoon ei päästä suoraa käsiksi. Tällöin ei voida vain uteliaisuudesta tarkastaa millaista avointa dataa on saatavilla ja pohtia millaisiin käyttötarkoituksiin kyseistä data-aineistoa olisi voisi mahdollisesti hyödyntää. Myös avoimien data-aineistojen löydettävyys sekä itse sivustoilta että hakukoneiden avulla koettiin käyttäjän kannalta haasteelliseksi. Tarjottua avointa dataa ei useinkaan ilmaistu organisaatioiden sivulla tarpeeksi selvästi käyttäjälle ja toisaalta tarjottu avoin data oli usein

nimetty puutteellisesti. Puutteellinen aineistojen nimeäminen voidaan nähdä haasteena silloin, kun sopivaa aineistoa ollaan kartoittamassa ja halutaan saada nopea katsaus tarjolla oleviin aineistoihin.

6. POHDINTA

Tutkimuksen tärkeimpinä havaintoina määriteltiin seitsemän erillistä avoimen datan hyödyntämisen haastetta, jotka näyttäytyvät erityisesti globaalin organisaation big data analytiikan näkökulmasta. Kaikki havaitut haasteet korostavat taustalla olevaa avoimen datan yleistä ongelmaa – avoimien datalähteiden yhdisteleminen on haastavaa ja työlästä eikä aineistojen sisältämää dataa kuvata tarpeeksi tarkalla tasolla. Aineistot eivät sellaisenaan ole useinkaan yhteensopivia toistensa kanssa ja eri aineistojen yhdisteleminen tuottaa paljon manuaalista työtä. Esimerkiksi dataformaattien yhtenäistäminen saattaa olla hyvinkin työlästä ja erityisesti yhdisteltäessä useita aineistoja havaintotarkkuus ja maantieteellinen kattavuus havaittiin yhdeksi suureksi haasteeksi kun aineiston tarkkuustasoa joudutaan väistämättä laskemaan. Toisaalta havaittiin myös, että aineistojen metatietojen ja laadun kuvaaminen oli lähes poikkeuksetta hyvin heikkoa tai vähintään se vaihteli erityisen paljon eri aineistojen välillä. Kaikki havaitut haasteet vaikeuttavat osaltaan avoimen datan hyödynnettävyyttä ja jatkokäyttöä. Tämän vuoksi analytiikan hyödyntäminen ja uusien löydösten tekeminen data-aineistoista vaikeutuu, tai vähintään siitä tulee työläämpää. Haasteet korostuvat erityisesti globaalin organisaation näkökulmasta, kun kattavan big data analytiikan mahdollistamiseksi joudutaan kartoittamaan useita eri datalähteitä ja yhdistelemään niiden sisältämää dataa, sillä avoin data havaittiin lähes poikkeuksetta maantieteellisesti rajoittuneeksi.

Työn tulosten perusteella korostui myös havainto siitä, että avoin data ei ole yhtäläisesti kaikkien käyttäjien hyödynnettävissä, vaikka se täyttäisi kaikki avoimelle datalle asetetut vaatimukset ja kriteerit. Avoimen datan hyödyntäminen osana globaalin organisaation big data analytiikkaa vaatii erityistä osaamista ja tuntemusta data-analytiikan aihepiiristä sekä eri ohjelmistojen tehokkaasta hyödyntämisestä, jonka vuoksi avoimen datan hyödyntäminen rajoittuu vain hyvin pieneen osaan käyttäjistä. Saman havainnon ovat tutkimuksissaan tehneet esimerkiksi Graves & Hendler (2013), Gurstein (2011) sekä Janssen et al. (2012), joiden mukaan teknisen asiantuntemuksen puute aiheuttaa eriarvoisuutta avoimen datan käyttäjien keskuudessa, eikä kaikilla käyttäjillä siten ole yhtäläistä mahdollisuutta hyötyä saatavilla olevasta avoimesta datasta. Myös Huijboom & Broek (2011) ovat EU-maiden avoimen datan strategioihin keskittyvässä selvityksessään todenneet, että saatavilla olevia avoimen datan tietokantoja tulisi muokata käyttäjäystävällisempään suuntaan, jotta kansalaiset ja yritykset voisivat hyötyä avoimesta datasta tehokkaammin.

Kynnystä avoimen datan hyödyntämiselle voidaan pyrkiä pienentämään parantamalla erityisesti aineiston metadatan kuvausta. Tällöin myös käyttäjät, jotka eivät omaa vahvaa tietoteknistä osaamista, voivat helpommin pystyä hyötymään avoimia data-

aineistoja kun ymmärrys aineiston sisällöstä paranee. Metadatan tärkeys on tunnistettu myös muissa avoimen datan hyödynnettävyyteen keskittyvissä tutkimuksissa, joista esimerkiksi Zuiderwijk et al. (2012) ovat todenneet puutteellisten metadatan kuvausten olevan yksi suurimmista avoimen datan uudelleenkäyttöä rajoittavista tekijöistä, kun käyttäjä ei ymmärrä saatavilla olevan aineiston sisältöä. Metadatan kuvauksessa tulisi erityisesti korostaa tietoja siitä, mitä millaista dataa aineisto pitää sisällään, missä muodossa se on saatavilla sekä mitä rajoitteita sen käyttöön ja keräämiseen liittyy. Kattava metadatan kuvaaminen parantaa siten erityisesti aineiston läpinäkyvyyttä ja luo luotettavuutta datan uudelleenhyödyntämisen näkökulmasta.

Kuten työn teoriaosuus sekä empiirinen osuus osoittavat, avoimelle datalle ei ole olemassa yhtenäistä määritelmää. Yhteisen määritelmän puute aiheuttaa haasteita myös sille, että avoin data ymmärretään monin paikoin eri tavalla ja avointa dataa on saatavilla hyvin erilaisessa muodossa sekä keskenään eriävillä kuvauksilla että käyttöehdoilla. Myös Borglund ja Engvall (2014) ovat tutkimuksessaan tehneet havainnon, että terminä avointa dataa käytetään hyvin laajasti eri asiayhteyksissä eikä käsite ole tästä johtuen vielä täysin vakiintunut. Usein tutkimukset määrittelevät itse avoimen datan käsitteen, jolloin osassa tutkimuksia avoin data nähdään avoimen hallinnon ja julkisten organisaatioiden kautta (Janssen et al. 2012; Borglund & Engvall 2014) tai osana avoimia innovaatioita ja data-analytiikkaa (Zuiderwijk et al. 2014; Manyikan et al. 2013).

Ratkaisuvaihtoehtona avoimen datan yhtenäistämiselle olisi erillisen avoimen datan standardin määrittäminen. Standardin tulisi pitää sisällään yksiselitteinen määritelmä avoimelle datalle sekä ehdot sille, miten avointa dataa tulisi tarjota ja avata uudelleenhyödynnettäväksi. Avoimen datan standardisoinnin puute nousee selkeästi esille vain muutamissa tutkimuksissa (Huijboom & Broek 2011, Janssen et al. 2012; Zuiderwijk et al. 2012), jonka vuoksi avoimelle datalle ei mitä luultavammin ole vielä pystytty määrittämään yhtenäistä määritelmää, joka olisi yleisesti käytössä. Nyt standardi on olemassa pelkästään avoimen datan käyttöehtojen määrittämiseen, mutta Creative Commons 4.0 -lisenssi ei ole vielä tutkimuksen havaintojen mukaan laajasti käytössä. Standardi ei siten ota mitenkään kantaa esimerkiksi avoimen datan yhtenäiseen dataformaattiin tai metadatan kuvaukseen.

Tutkimuksen havaintojen perusteella voidaankin todeta, että avoimen datan tuottajien tulisi huomioida entistä paremmin aineistojen loppukäyttäjät. Data tulisi avata sellaisessa muodossa, että dataa olisi mahdollisimman helppo hyödyntää uusiin käyttötarkoituksiin ja mahdollisesti uusiin analyyseihin, innovaatioihin ja liiketoimintamahdollisuuksiin. Mikäli avoimen datan hyödynnettävyyden useisiin haasteisiin ei kiinnitetä huomiota ja paranneta aineistojen uudelleenhyödyntämisen mahdollisuuksia, ei useinkaan päästä niihin tavoitteisiin, joiden vuoksi organisaatiot ovat alun perin alkaneet avointa dataa julkaisemaan. Huomion arvoista onkin, että avoimella datalla ei sellaisenaan ole arvoa, vaan sen arvo koostuu pelkästään uudelleenkäytön ja eri aineistojen yhdistelemisen lop-

putuloksena. Täten avoimien data-aineistojen loppukäyttäjien parempi huomioiminen ja aineistojen hyödynnettävyyden parantaminen olisi ensisijaisen tärkeää.

6.1 Tutkimuksen arviointi

Tutkimuksen ja sen toteutuksen arvioinnissa tulee kiinnittää huomiota useaan tekijään. Tutkimuksen tekijän onkin tärkeää pyrkiä arvioimaan omaa työtänsä kriittisesti ja pyrkiä arvioinnissa objektiivisuuteen. Tutkimuksen objektiivisuus muodostuu siitä, miten tulokset ovat löydettävissä tutkimuksen kontekstista eli ne eivät muodostu tutkijan omista näkökulmista tai mielipiteistä. (Soininen 1995, s. 122-148) Tutkimuksen arvioinnin objektiivisuus pyritään toteuttamaan siten, että arvioidaan työn luotettavuutta, rajoitteita sekä kontribuutiota olemassa oleviin tutkimuksiin aiheesta.

Tutkimuksen luotettavuuden arviointi on ongelmallista erityisesti kvalitatiivisen tutkimuksen yhteydessä (Soininen 1995, s. 122). Kvalitatiivinen tutkimus perustuu usein lähes pelkästään tutkijan omiin havaintoihin ja näkemyksiin, jonka vuoksi luotettavuuden määrittäminen on usein hyvin haastavaa. Kvalitatiivisen tutkimuksen luotettavuutta voidaan kuitenkin parantaa kertomalla tarkasti tutkimuksen kaikista vaiheista. Tässäkin tutkimuksessa aineiston valinta ja analysointi on pyritty kuvaamaan mahdollisimman yksityiskohtaisesti ja läpinäkyvästi sekä tutkimuksessa käytettyä avoimen datan hyödynnettävyyden mittaristoa ja sen valintoja on perusteltu myös teoriaan pohjautuen. Havainnot aineistosta pohjautuvat tutkijan omaan käsitykseen ja näkemykseen hyödynnettävyyden haasteita, mutta havaintojen tukena on pyritty hyödyntämään mahdollisimman paljon myös kirjallisuutta sekä aikaisempia tutkimuksia aiheesta, joka osaltaan tukee tutkimuksen tulosten luotettavuutta.

Tutkimuksen tulosten luotettavuuden arvioinnissa tarkastellaan yleensä kahta tekijää, reliaabeliutta ja validiutta. Reliabiliteetilla viitataan tutkimuksen mittaustulosten toistettavuuteen, eli siihen miten todennäköisesti myös toinen tutkija tai arvioida päätyy samaan lopputulokseen jos tutkimus toistetaan samoista lähtökohdista. Tutkimuksen validius puolestaan tarkastelee tutkimusmenetelmän tai asetetun mittarin kykyä mitata sitä, mitä tutkimuksen on tarkoituskin mitata. (Hirsjärvi et al. 2007, ss. 226-227) Tutkimuksen reliabiliteettia ja validiteettia pohdittaessa voidaan nostaa esille, että aineiston valinnalla on varmasti vaikutusta tutkimuksen lopputuloksiin, sillä erot eri avoimien data-aineistojen välillä olivat huomattavan suuria. Mikäli aineistoon valikoituu erittäin hyvin kuvatut ja laadukkaat avoimen datan lähteet korostuvat tutkimuksen tuloksissa varmasti eri asiat kuin heikommin kuvattujen lähteiden kohdalla. Aineiston valinta on kuitenkin pyritty tutkimuksessa kuvaamaan mahdollisimman läpinäkyvästi, jotta reliabiliteetti säilyy. Validiteetti on puolestaan pyritty varmistamaan sillä, että taustalla on hyödynnetty olemassa olevaa teoriaa aineistosta. On kuitenkin myös todettava, että tarkasteltava aihepiiri on hyvin uusi eikä parhaita käytäntöjä tai mittareita ole vielä kehitetty. Tämä lisää tutkimuksen uutuusarvoa, mutta validiteettia on haastava aukottomasti varmistaa ilman vastaavien tutkimusten tuloksia.

Työn rajoitteina on myös otettava huomioon tutkimusaineiston rajoittuvuus vain hyvin pieneen osaan avoimista datalähteistä sekä aihepiiriltään kapealle tarkastelualueelle. Tutkimusaineisto koostui 16 eri avoimesta datalähteestä, jotka keskittyvät lähes yksinomaan sää- ja paikkatietoa sekä tuuliturbiineista ja sähköverkoista saatavilla olevaan avoimeen dataan. Aihepiirien aiheuttama rajaus saattaa korostaa tiettyjä hyödynnettävyyden haasteita, kuten havaintotarkkuuden tai ajanjakson tarkastelu. Toisaalta avoimen datan hyödynnettävyydelle asetettu mittaristo ei välttämättä kata kaikkia mahdollisia kriteereitä avoimelle datalle, sillä mittaristo on määritelty erityisesti edellä mainittujen aihepiirien sisältämän avoimen datan tarkasteluun. Työn rajoitteista johtuen myös tutkimustulosten yleistettävyyttä on arvioitava kriittisesti. Kuten tutkimustuloksista huomattiin, aineistoista tehtyjen havaintojen eroavaisuudet olivat paikoittain huomattavia, jonka vuoksi tuloksissa korostuivatkin sellaiset avoimen datan hyödynnettävyyden haasteet, jotka esiintyvän aineistossa useasti tai jotka aiheuttivat merkittävää haittaa avoimen data-aineiston uudelleen hyödyntämiselle sekä yhdistämiselle muihin data aineistoihin osana globaalin organisaation big data analytiikkaa.

Pohdittaessa tutkimuksen kontribuutiota olemassa olevaan teoriaan, voidaan saatuja tuloksia vertailla Zuiderwijkin et al. (2012) tekemään tutkimukseen, joka on toteutettu kartoittamalla kirjallisuudesta löytyviä hyödynnettävyyden haasteita sekä keräämällä empiiristä aineistoa avoimen datan hyödyntämisestä haastatteluiden perusteella. He ovat tutkimuksessaan määrittäneet yleisimmät avoimen datan hyödynnettävyyteen liittyvät haasteet ja esteet, joita ovat saatavuus ja pääsy, löydettävyys, käytettävyys, ymmärrettävyys, laatu, yhdistettävyys, yhteensopivuus ja vertailukelpoisuus, metadata, tiedot datan tarjoajasta sekä avaaminen ja lataaminen. Voidaankin todeta, että tutkimusten tulokset ovat keskenään hyvin samankaltaisia, vaikkakin ne korostavat hieman eri teemoja keskenään. Tässä tutkimuksessa korostuivat yhdisteltävyyden, ymmärrettävyyden ja käytettävyyden haasteet, joita myös Zuiderwijk et al. (2012) ovat tunnistaneet. Ainut haaste, jota tässä tutkimuksessa ei havaittu oli tiedot datan tarjoajasta, jotka pääosin oli kuvattu riittävällä tarkkuudella tämän tutkimuksen aineistossa. Huomattavaa kuitenkin oli, että datan alkuperä ja arkkitehtuurikuvaus oli paikoittain hyvin hatarasti määritelty, joka voi osaltaan johtua puutteellisista tiedoista datan tarjoajassa.

6.2 Jatkotutkimusehdotukset

Niin julkisten kuin yksityistenkin organisaatioiden datan avaaminen on avannut uusia mahdollisuuksia liiketoiminnalle, sovelluskehitykselle sekä analytiikalle ja lisäarvopalveluille, mutta myös lisännyt läpinäkyvyyttä organisaatioiden toiminnassa. Kuten tässä tutkimuksessakin todettiin, avointa dataa on saatavilla yhä etenevässä määrin mutta sen uudelleenhyödyntäminen ja eri aineistojen yhdisteleminen on usein sekä haastavaa että työlästä, sillä yhteistä määritelmää tai standardia ei avoimelle datalle ole. Mielenkiintoista olisikin tulevaisuudessa tutkia, millaisiin tarkoituksiin avointa dataa pääasiassa hyödynnetään ja millaisessa muodossa se tulisi eri käyttäjille tarjota, jotta dataa voidaan

paremmin ja tehokkaammin hyödyntää osana organisaatioiden big data analytiikkaa. Jatkossa voitaisiin siis tutkia, millainen standardi avoimelle datalle voitaisiin asettaa että se palvelisi mahdollisimman montaa loppukäyttäjää ja helpottaisi aineistojen yhdistämistä sekä hyödyntämistä, mutta toisaalta ei myöskään aiheuttaisi datan tuottajalle liian suurta lisätyötä datan avaamiselle.

Tämän lisäksi olisi mielenkiintoista ajan myötä kartoittaa lisääntyykö avoimen datan määrä maailmassa niin eksponentiaalisesti kuin tutkimukset ovat olettaneet vaan aiheuttavatko tässäkin tutkimuksessa esille nousseet hyödynnettävyyden haasteet tilanteen, ettei avointa dataa nähdä niin houkuttelevana raaka-aineena vaan datasta ollaan ennemmin valmiita maksamaan tai hankkimaan muualta, jotta se saadaan loppukäyttäjälle oikeassa ja helposti hyödynnettävässä muodossa.

7. YHTEENVETO

Yhä useammat organisaatiot ovat alkaneet avaamaan dataansa kansalaisten, tutkijoiden ja muille organisaatioiden sekä sidosryhmien vapaasti saataville. (Zuiderwijk et al. 2012, s. 167) Vaikka avoimen datan ilmiö on vasta aluillaan, on sen nähty tuovan merkittävää lisäarvoa, kun edistyksellisen analytiikan avulla avoimia ja organisaation sisäisiä datalähteitä voidaan yhdistellä. (Manyikan et al. 2013, s. 2) Motivaationa datan avaamisen taustalla onkin usein tutkimuksen ja kehityksen vauhdittaminen sekä uusien innovaatioiden ja liiketoimintamahdollisuuksien syntyminen. (Janssen et al. 2012) Avoimelle datalle ei kuitenkaan ole vielä olemassa vakiintunutta määritelmää, joka osaltaan aiheuttaa haasteita ilmiöstä puhuttaessa. Usein avoimella datalla tarkoitetaan dataa, johon kuka tahansa voi avoimesti päästä käsiksi, käyttää, muokata ja jakaa sitä mihin tahansa käyttötarkoitukseen (Ubaldi 2013, s. 6). Datan tarjoajana voi toimia joko julkinen tai yksityinen organisaatio, ja se on saatavilla koneluettavassa muodossa ilman teknisiä rajoitteita ja erillisiä kustannuksia. (Open Knowledge Foundation 2015) Saatavilla oleva avoin data vaihtelee sää ja paikkatiedoista julkisen sektorin budjetteihin ja liikennetietoihin. (Janssen et al. 2012)

Big data on avoimen datan tavoin hyvin uusi ilmiö. (Gurin 2014b) Big datalla tarkoitetaan erityisen suurten ja järjestämättömien tietomassojen keräämistä, säilyttämistä ja ennen kaikkea analysointia tietoteknisten ratkaisujen avulla (Vakkuri 2013). Big data analytiikka nähdään siis työkaluna erityisen suurten datamäärien käsittelyyn, sillä sen avulla on mahdollista työstää suuria määriä sekä reaaliaikaista, että historiallista tietoa ja löytää uusia malleja tai poikkeavaisuuksia, jotka voivat osoittaa mahdollisuuksia uusille tuotteille ja palveluille tai tehokkaammille toimintatavoille. (Emanin et al. 2015, s. 71) Vaikka big data ja avoin data ovat käsitteinä täysin erillisiä, on ilmiöissä nähtävissä myös päällekkäisyyksiä ja yhdistäviä teemoja. Tässä tutkimuksessa keskityttiin hyvin suurten ja sisällöltään vaihtuvien avoimien data-aineistojen tarkasteluun, jolloin puhutaan avoimesta big datasta. (Marton et al. 2013, s. 2)

Tämän tutkimuksen motivaation taustalla on Tampereen teknillisen yliopiston tiedonhallinnan ja logistiikan laitoksen tutkimus, jossa tutkittiin ulkoisten datalähteiden saatavuutta ja hyödyntämismahdollisuuksia tuulivaihteita valmistavan globaalin yrityksen liiketoiminnassa. Tutkimuksessa kartoitettiin eri avoimen datan lähteitä ja niiden hyödyntämistä osana organisaation big data analytiikkaa, jolloin huomattiin useita eri hyödynnettävyyden haasteita avoimeen dataan liittyen. Havaituista haasteista johtuen tässä tutkimuksessa haluttiin keskittyä tarkemmin avoimen datan hyödynnettävyyden haasteiden tunnistamiseen. Tutkimusongelmana tällä tutkimukselle oli tietämättömyys siitä, mitä haasteita esiintyy avoimen datan hyödyntämisessä osana globaalin organisaation

big data analytiikkaa ja päätutkimuskysymykseksi asetettiin: ”Mitä haasteita esiintyy avoimen datan hyödyntämisessä globaalin organisaation big data analytiikassa?”. Tutkimuskysymykseen lähdettiin hakemaan vastausta empiirisen tutkimusmenetelmän avulla kartoittamalla saatavilla olevia avoimia datalähteitä erillisen mittariston avulla ja havainnoimalla avoimen datan hyödynnettävyyden haasteita. Tutkimus toteutettiin laadullisena dokumenttisanalyysinä, joka hyödyntää eksploratiivisen tutkimuksen strategiaa. Tutkimuksen taustalla on siten nähtävissä hermeneuttinen tieteenkäsitys, sillä tutkimuksessa korostuvat tutkijan oma tulkinta, käsitys sekä ymmärrys aiheesta.

Tutkimusprosessi jaoteltiin neljään erilliseen kokonaisuuteen. Ensimmäisenä kartoitettiin avoimia datalähteitä, jotka tutkimuksen rajausten mukaan sää-, paikkatieto-, tuuliturbiini- sekä sähköverkkodataan. Tutkimusaineistoksi valittiin 16 erillistä avointa datalähdettä. Tämän jälkeen jokainen tutkimusaineistoon kuulunut avoimen datan lähde arvioitiin määriteltyyn mittaristoon kuuluvien kymmenen erillisen arviointikriteerin perusteella. Arviointikriteerinä tarkasteltiin avoimen data-aineiston saatavuutta, kokonaisuutta ja laatua, dataformaatin avoimuutta ja koneluettavuutta, teknistä avoimuutta arkkitehtuurien ja rajapintojen avulla, käyttöehtoja ja uudelleenkäytettävyyden mahdollisuutta, kustannuksia, sekä datan ymmärrettävyyttä metadatan kuvaamisella. Tämän lisäksi tarkasteltiin data-aineiston maantieteellistä kattavuutta, havaintotarkkuutta sekä ajanjaksoa, jolta avoin data oli kerätty. Arvioinnin jälkeen datalähteiltä vertailtiin toisiinsa ja koostettiin havaintoja siitä, millaisia avoimen datan hyödynnettävyyden haasteita aineistossa esiintyi. Viimeisenä määriteltiin aineiston perusteella sellaiset avoimen datan hyödynnettävyyden haasteet, jotka nähtiin esiintyvän useasti tai jotka aiheuttivat merkittävää haittaa avoimen data-aineiston hyödyntämiselle.

Tutkimuksen keskeisinä havaintoina määriteltiin seitsemän haastetta, joiden nähtiin merkittävästi haittaavan tai hankaloittavan avoimen datan hyödyntämistä globaalin organisaation big data analytiikassa. Tutkimuksen tuloksena kartoitettiin seuraavat avoimen datan hyödynnettävyyden haasteet: tiedostomuotojen eroavaisuudet, puutteet metadatassa, erot havaintotarkkuuksissa, maantieteelliset rajoitteet, heikko arkkitehtuurikuvaus ja rajapinnat, eroavaisuudet datan laadussa sekä heikko saatavuus ja löydettävyys. Tutkimustuloksia pohdittaessa havaittiin myös, että suurin ongelma avoimen datan hyödyntämisen taustalla on, että avoimien datalähteiden yhdisteleminen on haastavaa ja työlästä. Aineistojen sisältämää dataa ei kuvata tarpeeksi tarkalla tasolla eivätkä aineistot yleisesti ottaen ole sellaisenaan yhteensopivia toisten aineistojen kanssa. Tämän vuoksi avoimen datan ei nähdä olevan yhtäläisesti kaikkien käyttäjien hyödynnettävissä, sillä avoimen datan hyödyntäminen osana globaalin organisaation big data analytiikkaa vaatii usein erityistä tietoteknistä osaamista ja tuntemusta data-analytiikan aihepiiristä sekä eri ohjelmistojen hyödyntämisestä. Saatuja tuloksia vertailla Zuiderwijkin et al. (2012) tekemään tutkimukseen, joka havaitsi pääosin hyvin samankaltaisia avoimen datan hyödynnettävyyden haasteita, vaikkakin tutkimukset korostavat keskenään hieman eri teemoja.

Ratkaisuna ongelmille voitaisiin asettaa erillinen avoimen datan standardi, joka määrittelisi yhteisesti avoimen datan käsitteen sekä ehdot sille, miten avointa dataa tulisi tarjota ja avata uudelleenhyödynnettäväksi. Tutkimuksen havaintojen mukaan nykytilassa avoimen datan loppukäyttäjää ei huomioida tarpeeksi. Data tulisi avata sellaisessa muodossa, että avointa dataa olisi mahdollisimman helppo hyödyntää uusiin käyttötarkoituksiin. Vain tällöin voidaan päästä avoimen datan taustalla oleviin tavoitteisiin, jotka korostavat uusien innovaatioiden, analyysien ja liiketoimintamahdollisuuksien kehittymistä. Jatkossa olisikin mielekästä tutkia, millainen standardi avoimelle datalla voitaisiin asettaa. Toisaalta ajan myötä olisi mielenkiintoista kartoittaa, lisääntyykö avoimen datan määrä maailmassa, vai aiheuttavatko tässäkin tutkimuksessa esille nousseet hyödyntävyyden haasteet tilanteen, jossa avointa dataa ei hyödynnetä ja nähdä niin houkuttelevana raaka-aineena uusien analyysien ja datasta tehtävien löydösten tekoon.

LÄHTEET

Ala-Pietilä, P. & Pennanen, R. (2013). 21 polkua kitkattomaan Suomeen. ICT 2015 – työryhmän raportti, Työ- ja elinkeinoministeriö. Saatavissa: http://www.tem.fi/files/35440/TEMjul_4_2013_web.pdf [Viitattu 10.9.2015]

Alasuutari, P. (1999). Laadullinen tutkimus. 3. uudistettu painos. Tampere, Vastapaino. 317 s.

Anttila, P. (1998). Tutkimisen taito ja tiedonhankinta. Saatavissa: http://www.metodix.com/fi/sisallys/01_menetelmat/01_tutkimusprosessi/02_tutkimisen_taito_ja_tiedon_hankinta/. [Viitattu 28.10.2015]

Apte, C.V., Hong, S.J., Natarajan, R., Pednault, E.P.D., Tipu, F.A. & Weiss, S.M. (2003). Data-intensive analytics for predictive modeling, IBM Journal of Research & Development. Vol. 47(1), pp. 17-23.

Barton, C. & Court, D. (2012). Making Advanced Analytics Work for You, Harvard Business Review. Saatavissa: https://hbr.org/2012/10/making-advanced-analytics-work-for-you/ar/1 [Viitattu 28.2.2016]

Bendini, I., Tankoyeu, I., Farazi, F., Leoni, D., Pane, J. & Leucci, S. (2014). Open Government Data: Fostering Innovation, Journal of Democracy. Vol.6(1), pp. 69-79.

Bennett, D. & Harvey, A. (2009). Publishing Open Government Data, W3C. Saatavissa: http://www.w3.org/TR/gov-data/ [Viitattu 21.8.2015]

Berners-Lee, T. (2006). Linked Data - Design Issues, W3. Saatavissa: http://www.w3.org/DesignIssues/LinkedData.html [Viitattu 19.8.2015]

Blakemore, M. & Craglia, M. (2006). Access to public-sector information in Europe: policy, rights and obligations, The Information Society. Vol.22(1), pp. 13-24.

Bose, R. (2009). Advanced analytics: opportunities and challenges, Industrial Management & Data Systems. Vol. 109(2), pp. 155 – 172.

Borglund, E. & Engvall, T. (2014). Open data? Data, information, document or record? Records Management Journal. Vol. 24(2), pp.163-180.

Chandarana, P. & Vijayalakshmi, M. (2014). Big Data Analytics Frameworks, International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), pp. 430-434.

Chen, H., Chiang, R. H. L. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact, MIS Quarterly. Vol.36(4), pp. 1-24.

Chui, M., Manyika, J., Groves, P., Farrell, D., Van Kuiken, S., Doshi, E. A. (2013). Open data: Unlocking innovation and performance with liquid information, McKinsey Global Institute, 103 p.

Conradie, P. & Choenni, S. (2014). On the barriers for local government releasing open data, Government Information Quarterly. Vol. 31, pp. 10-17.

Daintith, J. & Wright, E. (2008). A Dictionary of Computing (6 ed.), Oxford University Press. 608 p.

Davies, T. (2013). Open Data Barometer: 2013 Global Report. Saatavissa: http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf [Viitattu 10.9.2015]

Davies, K. & Patterson, D. (2012). Ethics of Big Data: Balancing Risk and Innovation, O'Reilly Media, 82 p.

Dawes, S. & Helbig, N. (2010). Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency, Electronic Government. Vol. 6228, pp. 50-60.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J. & Zijlstra, T. (2015). The Open Data Handbook. Saatavissa: http://opendatahandbook.org/guide/en/ [Viitattu 13.8.2015]

Emani, C. K., Cullot, N. & Nicolle, C. (2015). Understandable Big Data: A survey, Computer Science Review. Vol. 17, pp. 70-81.

Euroopan komissio. (2004). Proposal for a Directive of the European Parliament and of the Council establishing an infrastructure for spatial information in the Community (INSPIRE). Brussels: European Commission. Saatavissa: http://ec.europa.eu/smart-regulation/impact/ia_carried_out/docs/ia_2004/sec_2004_0980_en.pdf [Viitattu 19.8.2015]

Euroopan komissio. (2011). Avoin data: Innovoinnin, kasvun ja läpinäkyvän hallinnon moottori. Saatavissa: http://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=CELEX:52011DC0882&from=EN. [Viitattu 13.8.2015]

Euroopan komissio. (2013). Commission welcomes Parliament adoption of new EU Open Data rules, Euroopan komissio. Saatavissa: http://europa.eu/rapid/press-release_MEMO-13-555_en.htm [Viitatttu 3.9.2015]

Finlex. (2015). Henkilötietolaki 22.4.1999/523. Saatavissa: http://www.finlex.fi/fi/laki/ajantasa/1999/19990523 [Viitattu 14.8.2015]

Fioretti, M. (2011). Open Data: Emerging trends, issues and best practices, Laboratory of Economics and Management of Scuola Superiore Sant'Anna, Pisa. 34 p.

Fitzgerald, B. (2006). The Transformation of Open Source Software, Mis Quarterly. Vol. 30(3) pp. 587-598.

Franks, B. (2012). Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, Wiley, 336 p.

G8. (2013). Open data charter and technical index. Saatavissa: www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex [Viitattu 13.8.2015]

Graves, A. & Hendler, J. (2013). Visualization tools for open government data, Proceedings of the 14th Annual International Conference on Digital Government Research, Association for Computing Machinery, New York, NY, pp. 136-145.

Gurin, J. (2004a). Open Data Now. The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation, McGraw-Hill Education, 272 p.

Gurin, J. (2014b). Big data and open data: what's what and why does it matter?, The Guardian. Saatavissa: http://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government. [Viitattu 10.9.2015]

Gurstein, M.B. (2011). Open Data: Empowering the Empowered or Effective Data Use for Everyone?, First Monday. Vol.16(2).

Halonen, A. (2012). Näkökulmia avoimuuteen: Avoin data ja avoimuus Iso-Britannian hallinnossa, Suomen Lontoon Instituutti, 110 s.

Harisalo, R. (2008). Organisaatioteoriat, Tampere, Tampere University Press, 332 s.

Hirsjärvi, S., Remes, P. & Sajavaara, P. (2007). Tutki ja kirjoita. 13. painos, Helsinki, Kustannusosakeyhtiö Tammi, 448 s.

Hellberg, A-S. & Hedström, K. (2014). The story of the sixth myth of open data and open government, Transforming Government: People, Process and Policy. Vol.9(1), pp. 35-51.

Hoxha, J. & Brahaj, A. (2011). Open Government Data on the Web: A Semantic Approach, International Conference on Emerging Intelligent Data and Web Technologies, pp. 107-113.

Hjalmarsson, A., Johansson, N. & Rudmark, D. (2015). Mind the Gap: Exploring Stakeholders' Value with Open Data Assessment, IEEE Computer Society, pp. 1314-1323.

Huijboom, N. & Broek, T. V. D. (2011). Open Data: An International Comparison Of Strategies. European Journal Of Epractice, Nro. 12, pp. 4-16

Ishwarappa & Anuradha, J. (2015.) A Brief Intoruction on Big Data 5Vs Characteritics and Hadoop Technology, Procedia Computer Science. Vol. 48, pp. 319 – 324.

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. Government Information Quarterly. Vol. 28(4), pp. 446–456.

Janssen, K., Charalabidis, Y. & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government, Information Systems Management. Vol. 29, pp. 258–268.

Jin, X., Wah, B., Cheng, X. & Wang, Y. (2015). Significance and Challenges of Big Data Research, Big Data Research. Vol. 2, pp. 59–64.

Kambatla, K., Kollias, G., Kumar, V. & Grama, A. (2014.) Trends in big data analytics, Journal of Parallel and Distributed Computing. Vol. 74, pp. 2561-2573

Krishnan, K. (2013). Data Warehousing in the Age of Big Data, Morgan Kaufmann, 370 p.

Kuitunen, O., Poikola, A. & Kuikkaniemi, K. (2014). My Data – Johdatus ihmiskeskeiseen henkilötiedon hyödyntämiseen. Liikenne- ja viestintävirasto. Saatavissa: http://urn.fi/URN:ISBN:978-952-243-418-0 [Viitattu 10.9.2015]

Lausch, A., Schmidt,A. & Tischendorf, L. (2014). Data mining and linked open data – New perspectives for data analysis in environmental research. Ecological Modelling, pp. 5–17.

Lee, G. & Kwak, Y. H. (2012). An Open Government Maturity Model for social media-based public engagement, Government Information Quarterly. Vol. 29, pp. 492–503.

Lichtenthaler, U. (2011). Open Innovation, The Academy of Management Perspectives. Vol. 25(1), 2011, pp. 75-93.

Lindman, J., Rossi, M., & Tuunainen, V. K. (2013). Open Data Services: Research Agenda. In System Sciences (HICSS), 2013 46th Hawaii International Conference on System Science. IEEE, pp. 1239-1246.

Loshin, D. (2013). Big data analytics, From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph, Morgan Kaufmann. 142 p.

Maltby, D. (2011). Big data analytics, 74th Annual Meeting of the Association for Information Science and Technology (ASIST).

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S. & Doshi, E. A. (2013). Open data: Unlocking innovation and performance with liquid information, McKinsey Global Institute, 116 p.

Marton, A., Avital, M. & Jensen, T. (2013). Reframing Open Big Data, In Proceedings of the 21st European Conference of Information Systems.

Masip-Bruin, X., Ren. G-J., Serral-Gracià, R. & Yannuzzi, M. (2013). Unlocking the Value of Open Data with a Process-based Information Platform, IEEE International Conference on Business Informatics, pp. 331-337.

Meijer, A. & Thaens, M. (2009). Public Information Strategies: Making Government Information Available To Citizens. Information Polity. Vol. 14, pp. 31–45.

Niiniluoto, I. (1980). Johdatus tieteenfilosofiaan. Helsinki, Otava. 314 s.

Ohlhorst, F. J. (2012). Big Data Analytics: Turning Big Data into Big Money, Wiley, 176 p.

Ojasalo, K. & Moilanen, T. & Ritalahti, J. (2014). Kehittämistyön menetelmät: uudenlaista osaamista liiketoimintaan, Helsinki, Sanoma Pro Oy, 204 p.

Olkkonen, T. (1994). Johdatus teollisuustalouden tutkimustyöhön. 2.painos. Espoo, Teknillinen korkeakoulu. Raportti 152/1993/Teta. 143 s.

Open Knowledge Foundation. (2015). Avoimen tiedon määritelmä. Saatavissa: http://opendefinition.org [Viitattu 11.8.2015]

Peltola, V. (2014). Avointa dataa: Tämän takia ei ole avoimen datan startuppeja, Taloussanomat. Saatavilla: http://www.taloussanomat.fi/kumppaniblogit/2014/11/17/taman-takia-ei-ole-avoimen-datan-startuppeja/201415884/322 [Viitattu 25.8.2015]

Poikola, A., Kola, P. & Hintikka, K. (2010). Julkinen data - johdatus tietovarantojen avaamiseen, Liikenne- ja viestintäministeriö, Helsinki. 96s. Saatavilla: http://www.julkinendata.fi/ [Viitattu 24.8.2016]

Routio, P. (1990) Tuote ja tieto, Tiedon hakeminen teksteistä, Helsinki: Taideteollinen korkeakoulu, päivitetty 20.1.2005. Saatavissa: http://www2.uiah.fi/projects/metodi/040.htm [Viitattu 24.10.2015]

Russom, P. (2011). Big data analytics, TDWI Best Practices Report, Fourth Quarter. The Data Warehouse Institute (TDWI).

Saunders, M., Lewis, P. & Thornhill, A. (2009). Research methods for business students. 5. painos. Pearson Education. England, 614 p.

Shadbolt, N. & O'Hara, K. (2013). Linked Data in Government. IEEE Internet Computing. Vol.17(4), pp. 72-77.

Soininen, M. (1995). Tieteellisen tutkimuksen perusteet. Turku, Turun yliopiston täydennyskoulutuskeskus, 182 s.

Susha, I., Grönlund, Å. & Janssen, M. (2015). Organizational measures to stimulate user engagement with open data, Transforming Government: People, Process and Policy. Vol.9(2), pp. 181-206

Tammisto, Y. & Lindman, J. (2012). Definition of Open Data Services in Software Business, Third International Conference on Software Business, pp. 297–303.

Tene, O. & Polonetsky, J. (2012). Privacy in the Age of Big Data: A Time for Big Decisions, Stanford Law Review, pp. 64-69.

Tuomi, J. (2007). Tutki ja lue: Johdatus tieteellisen tekstin ymmärtämiseen. Helsinki, Tammi. 171 s.

Tuomi, J. & Sarajärvi, A. (2009). Laadullinen tutkimus ja sisällönanalyysi. Helsinki, Tammi. 175 s.

The Economist. (2013). Open data: a new goldmine, Saatavissa: www.economist.com/news/business/21578084-making-official-data-public-could-spur-lots-innovation-new-goldmine [Viitattu 13.8.2015]

Ubaldi, B. (2013). Open Government Data: Ttowards Empirical Analysis of Open Government Data Initiatives, OECD Working Papers on Public Governance, Nro. 22, OECD, Saatavissa: http://dx.doi.org/10.1787/5k46bj4f03s7-en

Vajjhala, N. R., Strang, K. D. & Sun, Z. (2015.) Statistical Modeling and Visualizing Open Big Data Using a Terrorism Case Study, IEEE 3rd International Conference on Future Internet of Things and Cloud, pp. 489-496.

Vakkuri, M. (2013). Big data muuttaa maailmaa. Talouselämä. Saatavissa: http://www.talouselama.fi/kumppaniblogit/tieto/big+data+muuttaa+maailmaa/a2191461 [Viitattu 10.9.2015]

Willinsky, J. (2006). The Access Principle: The Case for Open Access to Research and Scholarship, Massachusetts Institute of Technology, 287 p.

Yhdysvaltojen presidentin kansila. (2009). Office of Management and Budget, Executive Office of the President, Open Government Directive. Saatavissa: https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf [Viitattu 19.8.2015]

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R. & Alibaks, R. S. (2012). ocio-technical Impediments of Open Data" Electronic Journal of e-Government. Vol. 10(2), pp. 156 – 172.

Zuiderwijk, A., Helbig, N., Gil-García, J. R. & Janssen, M. (2014). Special Issue on Innovation through Open Data - A Review of the State-of-the-Art and an Emerging Research Agenda: Guest Editors' Introduction, Journal of Theoretical and Applied Electronic Commerce Research. Vol. 9(2), 13 p.

Zhang, J., Dawes, S. & Sarkis, J. (2005). Exploring Stakeholders' Expectations Of The Benefits And Barriers Of E-Government Knowledge Sharing. Journal Of Enterprise Information Management, Vol.18(5), pp. 548-567.

LIITE A: TUTKIMUSAINEISTON KUVAUS

1. Weather databases

1.1 NCEI - National Centers for Environmental Information

Examination	Info
Administrator:	National Oceanic and Atmospheric Administration (NOAA)
Data:	Weather and climate data. For example land-based, marine, model, radar, weather balloon, satellite, and paleoclimatic datasets
Variables:	Land-based data: Temperature, dew point, relative humidity, precipitation, wind speed and direction, visibility, atmospheric pressure, and types of weather occurrences such as hail, fog, and thunder
Observation accuracy:	Sub-hourly, hourly, daily, monthly, annual, and multiyear timescales
Area:	Global, most detailed data from the USA
Time period:	1800s for certain data types and locations - present.
Availability:	Open data
Information architecture:	Data sources from multiple partnerships
Interface:	JSON v2 API
Data quality:	NOAA information quality guidelines
Costs:	Online access to most data is free
Data format(s):	ASCII, ArcGIS, KMZ, PDF, CSV
Metadata:	Metadata and content of the dataset described briefly, also historical observing metadata repository available
Terms of use:	Registration required for API-key
Access:	NCEI Data Access , offers also National Weather Service for the USA

1.2 European Climate Assessment & Dataset

Examination	Info
Administrator:	ECA&D Project Team, Royal Netherlands Meteorological Institute (KNMI)
Data:	Climate data from each station 75 indices are calculated, each describing changes in the mean or extremes of climate.
Variables:	Cloudiness, Cold, Compound, Drought, Heat, Humidity, Pressure, Rain, Snow, Sunshine, Temperature
Observation accuracy:	Daily, Monthly, Yearly, the winter-half (ONDJFM), the summer-half (AMJJAS), winter (DJF), spring (MAM), summer (JJA), autumn (SON)
Time period:	1918 – present
Area:	Europe
Availability:	Open data, Registration required
Information architecture:	The dataset consists of daily station series obtained from climatological divisions of National Meteorological and Hydrological Services and station series maintained by observatories and research centers throughout Europe and the Mediterranean. For more detailed info: Participant list
Interface:	No API interface
Data quality:	Quality control procedures are applied to all series using various algorithms. Detailed information: ATBD
Costs:	Free
Data format(s):	ASCII
Metadata:	No separate metadata description, only a brief explanation of notations
Terms of use:	Freely available for non-commercial research and education
Access:	ECA&D Daily data

1.3 Finnish Meteorological Institute

Examination	Info
Administrator:	Finnish Meteorological Institute
Data:	Weather, marine and climate observations. Real-time observations, time series and forecasts. Most relevant database: Weather observations
Variables:	Air temperature, Relative humidity, Atmospheric pressure, Wind, Rainfall, Total cloud cover and cloud height, Visibility, Snow depth, Solar radiation
Observation accuracy:	Every 10min from 2010 onwards, at least daily
Area:	Finland
Time period:	1959 – present
Availability:	Open data, Registration required
Information architecture:	No mention
Interface:	WFS 2.0, API-key needed, more info in Open data manual
Data quality:	Data is certified with ISO 9001:2008 standard. Weather observation data currently consists of mostly automatic weather stations, but some of the data is also gathered manually.
Costs:	Free
Data format(s):	XML
Metadata:	Clear and comprehensive description of the metadata
Terms of use:	Users must be approved by the Finnish Meteorological Institute open data license
Access:	The Finnish Meteorological Institute's open data

1.4 NSIDC - National Snow and Ice Data Center

Examination	Info
Administrator:	National Snow and Ice Data
Data:	Scientific data sets, focusing on the cryosphere and its interactions. Data is from satellites and field observations, NASA, NSF, NOAA, and other programs
Variables:	GPS coordinates, time period, keyword search
Observation accuracy:	Multiple
Area:	Global
Time period:	1800s for certain data types and locations - present.
Availability:	Open data
Information architecture:	Sponsored Programs are Antarctic Glaciological Data Center, Exchange for Local Observations and Knowledge of the Arctic (ELOKA), NOAA at NSIDC and NASA Distributed Active Archive Center at NSIDC (NSIDC DAAC)
Interface:	WMS, WFS, WCS, OpenSearch API: Data Set and Granule, SubsetAgent/GetData. More info in API guide
Data quality:	Quality and reliability of the data is guaranteed by a systematic approach to scientific stewardship of the data. More detailed: Data policies and standards
Costs:	Free
Data format(s):	ASCII, ArcGIS Native, Binary, CSV, Documents, GeoTIFF, HDF/HDF-EOS,HTML, KML/GML, MATLAB, XLS, NetCDF, etc.
Metadata:	Only temporal coverage, parameter, data format and summary
Terms of use:	Registration required, the data set source should be properly cited when the data are used.
Access:	Data at NSIDC

1.5 OpenWeatherMap

Examination	Info
Administrator:	OpenWeatherMap
Data:	Current, 5 day forecast and weather maps
Variables:	Temperature, Wind, Cloudiness, Pressure, Humidity, Sunrise, Sunset
Observation accuracy:	Updated every hour
Area:	Global
Time period:	Present and forecast, historical data available only at a charge
Availability:	Most open data
Information architecture:	NOAA, Environment Canada, European Centre for Medium-Range Weather Forecasts (ECMWF), Japan Meteorological Agency, METAR data from airports, APRS network and more than 40,000 private weather stations
Interface:	Only API interface
Data quality:	The quality of the data is not specified
Costs:	Free for basic data, more comprehensive data sets available at a charge
Data format(s):	JSON, XML, HTML
Metadata:	Metadata not defined
Terms of use:	Registration required, see more detailed terms
Access:	OpenWeatherMap

1.6 Weather Underground

Examination	Info
Administrator:	Weather Underground, Inc
Data:	Current, forecast and historical weather data, maps and radar data
Variables:	Temperature, Dew Point, Humidity, Pressure, Visibility, Wind Direction, Wind Speed, Gust Speed, Precipitation, Events, Conditions
Observation accuracy:	Forecast, Current, Historical data: daily, weekly, monthly and custom
Area:	Global
Time period:	1945 - present
Availability:	Open data
Information architecture:	100,000+ members sending real-time data, NDFD
Interface:	JSON, XML API-key needed
Data quality:	BestForecast -system, more detail info
Costs:	Data free when searched manually, API key available at a charge
Data format(s):	Comma Delimited Text File
Metadata:	Metadata not defined
Terms of use:	Data may only be used by you for personal, non-commercial purposes. If you want to use that data for commercial purposes, contact the company.
Access:	Historical weather

1.7 Merra

Examination	Info
Administrator:	NASA - National Aeronautics and Space Administration Goddard Earth Sciences (GES), Data and Information Services Center (DISC)
Data:	Atmospheric, land and ocean observations from satellites, aircrafts, ships, and other sources, More information on MERRA Data Products
Variables:	Multiple variables for atmospheric and climate data, keyword search
Observation accuracy:	Daily, Monthly, Diurnal
Area:	Global
Time period:	1979 – present
Availability:	Open data
Information architecture:	Datasets from Alaska Satellite Facility SAR Data Center (ASF SDC), NASA Langley Research Center Atmospheric Science Data Center (LaRC ASDC), Crustal Dynamics Data Information System (CDDIS), National Snow and Ice Data Center (NSIDC), Global Hydrology Resource Center (GHRC), Oak Ridge National Laboratory (ORNL), Goddard Earth Sciences Data and Information Services Center (GES DISC), Ocean Biology Processing Group (OBPG), Land Processes (LP), Physical Oceanography (PO), Level 1 Atmosphere Archive and Distribution System (MODAPS LAADS), Socioeconomic Data and Applications Data Center (SEDAC)
Interface:	No API interface, see Data services
Data quality:	Data are being uploaded to the MDISC after undergoing quality assurance in the GMAO. More detail info: MERRA Products
Costs:	Free
Data format(s):	HDF, NetCDF
Metadata:	Short description of the metadata, essential information missing
Terms of use:	The data set source should be properly cited when the data are used
Access:	Merra Data Services

1.8 Data.gov Climate

Examination	Info
Administrator:	U.S. General Services Administration, Office of Citizen Services and Innovative Technologies
Data:	Open data
Variables:	Multiple variables for climate: Coastal flooding, food resilience, water, ecosystem vulnerability, human health, energy infrastructure, transportation and the Arctic region.
Observation accuracy:	Multiple
Area:	North America
Time period:	1800 - present
Availability:	Open data
Information architecture:	Large amount of data sources
Interface:	No API interface
Data quality:	Data accessed through the Information Quality Act (P.L. 106-554)
Costs:	Free
Data format(s):	Multiple
Metadata:	Coherent metadata description for every data set, see more detailed standards, ISO-19139 Metadata
Terms of use:	No mention. Recommended to refer to the original source.
Access:	Data.gov Climate

2. Geographic information databases

2.1 NLS - National Land Survey of Finland

Examination	Info
Administrator:	National Land Survey of Finland
Data:	Orthophotos, elevation models and topographic database and various maps
Variables:	Background map series, Basic map raster 1:20 000. Benchmarks, Elevation model 2 m and 10 m, Elevation zones raster, General map 1:1 000 000, 1:4 500 000, General map raster 1:1 000 000, 1:2 000 000, 1:4 500 000, 1:8 000 000, Laser scanning data, Municipal Division in Finland, NLS aerial photographs, NLS orthophotos, Place names, Road names, Shaded relief raster, Topographic map raster 1:50 000, Topographic map 1:100 000 and 1:250 000, Topographic map raster 1:100 000. 1:250 000, 1:500 000, The Topographic database
Observation accuracy:	Datasets are updated at least once a year, some weekly
Area:	Finland
Time period:	Only present data available
Availability:	Open data
Information architecture:	No mention
Interface:	WMS, WFS, WCS, API-key needed
Data quality:	The quality of the data is specifed for every dataset.
Costs:	Free. For large bodies of data, ordering is a better option than downloading the dataset one piece at a time from the file service. Detachment and handling fees will be charged from orders.
Data format(s):	Multiple, e.g. LAZ, TIFF, GML, shape, mif, PNG, ASCII, XYZ, GML, txt. The Data formats cannot be choose, depends on the dataset and it content
Metadata:	Clear and comprehensive description of the metadata
Terms of use:	The data is available to public and companies freely and free of charge. The open data products can be used without compensation and with extensive and permanent rights of use.
Access:	NLS Open data

2.2 EuroGeographics

Examination	Info
Administrator:	Membership association and acknowledged voice of the European National Mapping, Cadastre and Land Registry Authorities.
Data:	EuroGlobalMap: open data 1:1 million scale topographic dataset covering 45 countries and territories in the European region
Variables:	Administrative boundaries, The water network, Transport networks, Settlements, Elevation, Names locations
Observation accuracy:	Released in January 2014
Time period:	Present
Area:	Europe
Availability:	Open data
Information architecture:	No mention
Interface:	No API interface
Data quality:	The metadata is based on the ISO 19115 standard. More info in Full product specification
Costs:	Free
Data format(s):	Geodatabase, Shapefile
Metadata:	Clear and comprehensive description of the metadata
Terms of use:	EuroGeographics open data license needed. It may be used for any legal purpose, including commercial exploitation. The intellectual rights and the source of the data are mandatory to acknowledge.
Access:	EuroGlobalMap

2.3 Natural Earth

Examination	Info
Administrator:	A collaboration of many volunteers and is supported by North American Cartographic Information Society (NACIS)
Data:	Public domain map dataset at 1:10m, 1:50m, and 1:110 million scales. Tightly integrated vector and raster data. Cultural, Physical, and Raster categories.
Variables:	Countries, Details, Boundary Lines, Breakaway, Disputed Areas, States, Provinces, Populated Places, Roads, Railroads, Airports, Ports, Urban Areas, Parks and Protected Lands, Time zones, Cultural Building Blocks and more
Observation accuracy:	Released in 2009, last update in January 2015
Time period:	Present
Area:	Global
Availability:	Open data
Information architecture:	No mention
Interface:	No API interface
Data quality:	The quality of the data is not specified
Costs:	Free
Data format(s):	Vector data: ESRI shapefile format, Raster data: TIFF format with a TFW world file
Metadata:	Metadata not defined, short description of the dataset content and version history
Terms of use:	The data can be used in any manner, including modifying the content and design, electronic dissemination, and offset printing for personal, educational, and commercial purposes.
Access:	Natural Earth Downloads

2.4 GeoPlatform

Examination	Info
Administrator:	GeoPlatform.gov. Developed by the member agencies of the Federal Geographic Data Committee (FGDC) through collaboration with partners and stakeholders.
Data:	Geographic data, imagery, applications, documents, web sites and other resources
Variables:	Multiple, keyword search
Observation accuracy:	Multiple, updated at least monthly
Time period:	Multiple
Area:	USA
Availability:	Open data
Information architecture:	Data source and maintainer is Data.gov. GeoPlatform.gov searches and only returns geospatial data
Interface:	CSW API, more information
Data quality:	All information accessed through GeoPlatform.gov is subject to the Information Quality Act (P.L. 106-554). For all data accessed through GeoPlatform.gov, each agency has confirmed that the data being provided through this site meets the agency's Information Quality Guidelines
Costs:	Free
Data format(s):	Multiple
Metadata:	Coherent metadata description for every data set, ISO-19139 Metadata
Terms of use:	-
Access:	GeoPlatform Dataset Catalog

3. Electricity grid and turbine databases

3.1 Entso-E

Examination	Info
Administrator:	European Network of Transmission System Operators for Electricity
Data:	Production, Consumption and Exchange of Electricity, EIC codes Other miscellaneous data: Net generating capacity of a specific year, Parallel load of each country, Lengths of circuits of all the countries, Transformers of all the countries, Inventory of generation of all countries
Variables:	Production, Consumption, Exchange (also specific for wind power)
Observation accuracy:	Monthly, Yearly
Area:	Europe
Time period:	2009 – present
Availability:	Open data
Information architecture:	No mention
Interface:	No API interface
Data quality:	Guidelines for Monthly Statistics Data Collection
Costs:	Free
Data format(s):	HTML, XLS
Metadata:	Metadata not defined
Terms of use:	The use of data without explicitly mentioning the source and indicating ENTSO-E the usage of its data is strictly prohibited.
Access:	Entso-E Database

3.2 Open Energy Information

Examination	Info
Administrator:	Wiki. Sponsored by U.S. Department of Energy, National Renewable Energy Laboratory, Reegle
Data:	Buildings, Geothermal, Hydrogen, Smart grid, Solar, Utilities, Water and Wind data
Variables:	Multiple
Observation accuracy:	Multiple
Area:	Global, most detailed data from the USA
Time period:	Multiple
Availability:	Open data
Information architecture:	Different datasets from multiple sources
Interface:	OpenEI API (REST)
Data quality:	All users can view, edit and add data. The source of the data is determined.
Costs:	Free
Data format(s):	CSV, XLS, ZIP, PDF, txt, etc.
Metadata:	Metadata not defined, depending of the dataset some additional info descripted
Terms of use:	No terms of use mentioned
Access:	OpenEI datasets

3.3 U.S. Department of Energy

Examination	Info
Administrator:	U.S. Department of Energy, Office of Electricity Delivery & Energy Reliability
Data:	Electric Disturbance Events (OE-417) Annual Summaries, Power outages. Data presents what government agencies have reported to the Department of Energy about grid outages in their region or sector.
Variables:	Month, Date event began, Time event began, Date of restoration, Time of restoration, Respondent, Area Affected, NERC Region, Alert Criteria, Event Type, Demand Loss (MW), Number of Customers Affected
Observation accuracy:	To the minute, Yearly summaries
Area:	USA
Time period:	2000 - present
Availability:	Open data
Information architecture:	No mention
Interface:	No API interface
Data quality:	Department of Energy Information Quality Guidelines Only major electricity providers and operators are required to report outages, so this database is not comprehensive
Costs:	Free
Data format(s):	XLS, PDF
Metadata:	Metadata not defined
Terms of use:	-
Access:	U.S Annual Electric Disturbance Events

3.4 Global Atlas for Renewable Energy

Examination	Info
Administrator:	International Renewable Energy Agency (IRENA)
Data:	Multiple datasets available on wind, solar, marine, geothermal and bio-energy.
Variables:	The Atlas enables to overlay maps of resources, protected areas, grids (available in some places), slope and landcover. It is possible to prospect sites of interest anywhere in the world using a large library of datasets called the data catalogue. It is also possible to create and save your own version of the Atlas, with the datasets you find most interesting, centered over your area of interest. Catalog is the dataset library of the Global Atlas and all the information in the Catalog can be accessed directly from the Global Atlas GIS interface. It enables the user to overlay information listed in a catalog of more than 1,000 datasets, and to identify areas of interest for further prospection.
Observation accuracy:	No mention
Area:	Global
Time period:	Time defined for every dataset individually. Most data available from early 2010s.
Availability:	Open data
Information architecture:	67 countries and more than 50 institutes and partners were contributing to the initiative.
Interface:	No API interface, mobile application available.
Data quality:	No mention
Costs:	Free
Data format(s):	XLS
Metadata:	Specific metadata description for each dataset.
Terms of use:	Registration required to certain information and certain functionality of the Global Atlas. See more detailed info.
Access:	Global Atlas for Renewable Energy