Tampereen yliopisto

Aleksi Oja

AVOIMEN TIEDON METATIETOJEN HAASTEET AVOIMIEN AINEISTOJEN HYÖDYNTÄMISESSÄ

Diplomityö
Tekniikan ja luonnontieteiden tiedekunta
Tarkastaja: Pekkola Samuli
Tarkastaja: Ilvonen Ilona
Toukokuu 2021

TIIVISTELMÄ

Aleksi Oja: Avoimen tiedon metatietojen haasteet avoimien aineistojen hyödyntämisessä
Diplomityö
Tampereen yliopisto
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
2021

Tämän laadullisen tapaustutkimuksen taustalla on Valtiovarainministeriön aloittama *Tiedon hyödyntäminen ja avaaminen* -hanke. Hankkeen tavoitteena on parantaa hyödynnettävän ja avattavan tiedon laatua ja yhteentoimivuutta teknisesti sekä semanttisesti. Yhtenä osatavoitteena on tunnistaa ja dokumentoida metatietoihin kohdistuvat muutostarpeet. Aineistojen metatietoja on tutkittu runsaasti tieteessä. Kuitenkin yksilöllisiä metatietojen haasteita ei ole tunnistettu avoimen tiedon hyödyntämisessä Suomessa. Tässä tutkimuksessa olevan teemahaastattelut ja tulokset luovat uutta arvokasta tietoa avoimen tiedon kehityskohteista.

Tässä tutkimuksessa esitetään 17 tunnistettua avoimen tiedon metatietojen haastetta ja huomioon otettavaa asiaa. Haasteet ja huomiot ovat ketjuutuneet ja vaikuttavat toisiinsa syy-seuraussuhteilla. Näitä suhteita kuvataan havainnollistavalla verkolla, jossa on kuvattu haasteet ja niiden vaikutukset toisiinsa. Haastatteluiden vastausten perusteella tunnistettujen haasteiden ja huomioiden syy-seuraussuhteet kohdistuvat neljään huomioon, jotka ovat metatietojen heikko laatu, aineiston heikko hakukonelöydettävyys, aineiston heikko luotettavuus ja aineiston heikko hyödynnettävyys. Näitä tunnistettuja haasteita on myös huomioitu muissa tieteellisissä tutkimuksissa.

Kolme merkittävää tämän tutkimuksen huomiota ovat edellä mainitut merkittävimmät avoimen tiedon metatietojen haasteet, haasteiden semanttinen luonne teknisyyden sijaan sekä haasteiden ketjuuntuneisuus ja suhteet toisiinsa.

Avainsanat: Avoin tieto, metatieto, PSI-direktiivi, yhteentoimivuus, semanttisuus

ABSTRACT

Aleksi Oja: Metadata challenges in the utilization of open data
Master of Science Thesis
Tampere University
Master's Degree Programme in Information & Knowledge Management
2021

The background to this qualitative case study is the *Tiedon hyödyntäminen ja avaaminen* -project (Utilization & Opening of Open data -project) launched by the Ministry of Finance in Finland. The aim of the project is to improve the quality & interoperability of open data technically & semantically. One part of the goal is to identify & document the required changes to metadata. The metadata of open data has been moderately studied in science. However, individual metadata challenges have not been identified in the utilization of open data in Finland. The semi structured interviews & results of this study create new valuable information about the development targets of open data.

This study presents 17 identified challenges to open data's metadata & issues to consider. Challenges & considerations are chained & interact with cause-and-effect relationships. These relationships are described by an illustrative network that describes the challenges & their implications for each other. The cause-and-effect relationships of the challenges & considerations identified on the basis of the responses to the interviews focus on four considerations: poor quality of metadata, poor search engine findability of the open datasets, poor reliability of the open datasets, & poor usability of the open datasets. These identified challenges have also been identified in other scientific studies.

The three major highlights of this study are the major challenges of open data metadata mentioned above, the semantic nature of the challenges rather than the technicalities, & the links & interrelationships of the challenges.

Keywords: Open data, metadata, PSI-directive, interoperability, semantic

ALKUSANAT

Haluan kiittää *Tiedon hyödyntäminen ja avaaminen* -hanketta ja hankkeen työryhmää tämän tutkimuksen aihepiirin mahdollistamisesta ja tutkimuksen tukemisesta. Suuri kiitos myös tähän tutkimukseen osallistuville haastateltaville henkilöille ja organisaatioille. Kiitän myös CGI Suomi Oy:tä taustatuesta. Kiitos kuuluu myös Tampereen yliopistolle ja tutkimuksen ohjaajilleni. Kiitän lisäksi perhettäni ja opiskelijakollegojani opintojeni tukemisesta koko yliopisto-opintojeni ajan.

Tämä tutkimus sai minun mielenkiintoni kohoamaan avointa tietoa ja sen hyödyntämistä kohtaan. Lisäksi ymmärrykseni metatiedoista kasvoi tutkimuksen aikana tuntuvasti. Metatietoa on olemassa kaikkialla, ja koen sen ymmärtämisen auttavan minua tulevaisuuden tehtävissäni. Huomasin kuitenkin, että tämä oli vasta ensiraapaisu avoimeen tietoon sekä sen metatietoihin, ja vielä on paljon opittavaa aiheesta.

Tampere, 16.05.2021

Aleksi Oja

SISÄLLYSLUETTELO

KUVALUETTELO

LYHENTEET JA MERKINNÄT

API	Ohjelmointirajapinta, engl. Application Programming Interface
CSV	Yleisesti käytetty taulukkomuotoinen tiedostomuoto, engl. Comma-seperated values
DCAT	RDF-viitekehyksen sanasto, engl. Data Catalog Vocabulary
GDPR	Yleinen tietosuoja-asetus, engl. General Data Protection Regulation
GSIM	Yleinen tilastotietomalli, engl. Generic Statistical Information Model.
JSON	Avoimen standardin tiedostomuoto, engl. JavaScript Object Notation.
NISO	National Information Standards Organization.
RDF	Standardoitu viitekehys avoimelle tiedolle engl. Resource Description Framework
URI	Tiedon paikan osoittama merkkijono, engl. (Uniform Resource Identifier).
W3C	World Wide Web Consortium
XML	Merkintäkielien standardi, engl. eXtensible Markup Language.

1. JOHDANTO

Tämä diplomityö tehdään valtiovarainministeriön aloittaman *Tiedon hyödyntäminen ja avaaminen* -hankeen tueksi luomalla uutta ymmärrystä tietojen avaamisen ongelmista metatietojen näkökulmasta. Hanke aloitettiin vuoden 2020 alussa ja siinä on mukana useita julkisia ja julkisomisteisia toimijoita Suomesta. Sen tavoitteena on kehittää tiedon avoimuutta ja hyödynnettävyyttä kansallisesti sekä edistää julkisen tiedon entistä laajempaa ja tehokkaampaa hyödyntämistä koko yhteiskunnassa. Näin hanke tukee Suomen kansallisen tietopolitiikan kantavia periaatteita, edesauttaa uuden tiedon luomista, sekä mahdollistaa valtavan määrän uusia kehitys- ja liiketoimintamahdollisuuksia lukuisille eri julkisille ja yksityisille toimijoille (Valtiovarainministeriö 2020).

Hankkeen on määrä valmistua vuoden 2022 lopussa, johon mennessä on löydetty ja ratkaistu uusia tiedon hyödyntämisen haasteita, luotu uusia toimintamalleja tiedon avaamiseen sekä luotu kansalliset API-linjaukset tiedon omistajille tiedon jakamista varten. Näihin ja muihin tavoitteisiin päästään neljän työpaketin avulla. Näitä työpaketteja johtaa hankkeen vetäjänä toimiva hanketyöryhmä, joka koostuu useista hankkeessa olevien organisaatioiden toimijoista. Tämä diplomityö tukee näitä työpaketteja ja hankkeen tavoitteita tiedon semanttisesta ja teknisestä yhteentoimivuudesta metatietojen näkökulmasta. Yhtenä hankkeen osatavoitteena on tunnistaa ja dokumentoida metatietoihin kohdistuvat muutostarpeet. Tällä diplomityöllä pyritään edesauttamaan tätä tavoitetta, sekä luomaan yleisesti uutta tietoa metatietojen haasteista avoimessa tiedossa.

Avoin tieto on levinnyt yleiseen tietoisuuteen globaalisti, ja sen merkitys on kasvanut lähivuosikymmeninä valtavasti. Maailmanlaajuinen paine aineistojen avaamiselle julkista käyttöä varten on kohdistunut julkisiin organisaatioihin (Attard *et al.* 2015). Suomessa avointa tietoa alettiin hyödyntämään enenevissä määrin vuoden 2009 jälkeen, kun Euroopan unioni julkaisi avointa tietoa koskevan PSI-direktiivin, jota on uudistettu sen jälkeen useamman kerran. Vastikään uudistuneen PSI-direktiivin ja Suomen tietopoliittisten linjausten seurauksena avoimen tiedon merkitys kasvaa kansallisesti entisestään lähivuosina. Suomen eri julkiset organisaatiot ovat jo kauan avanneet tietoaineistojaan koko yhteiskunnan käyttöön.

Eri aineistoja avattaessa ja hyödynnettäessä voi tulla vastaan useita erilaisia ongelmia metatietojen suhteen. Tietovarantojen metatietojen oikeellisuus voi heiketä, tai sen merkitys voi muuttua, kun tietovarannon tiedot siirtyvät eri rajapintojen läpi. Metatiedoilla on myös suuri merkitys aineiston ymmärrettävyyteen, hyödynnettävyyteen ja luotettavuuteen (Kubler *et al.* 2018). Aineistojen metatiedoille on olemassa paljon viitekehyksiä ja standardeja, jotka selkeyttävät ja tukevat metatietojen laadun ylläpitämistä sekä helpottavat aineistojen käyttöä. Tästä huolimatta avoimissa aineistoissa on haasteita, jotka rajoittavat aineistojen hyödynnettävyyttä sekä metatietojen teknistä ja semanttista yhteentoimivuutta.

Metatietoa on tutkittu tieteellisesti monesta eri näkökulmasta hyvin kauan ajan aina Aristoteleesta asti (Pomerantz 2015, s. 5). Etenkin metatiedon laatua ja sen käyttämistä eri tiedostomuotojen, standardien ja viitekehysten kanssa on tutkittu paljon (Aroyo *et al.* 2011; Attard et al. 2016). Avoimen tiedon näkökulmasta metatietojen haasteita on tutkittu vasta hieman (Kubler *et al.* 2018). Kuitenkin haasteiden kattavaa tunnistamista ja määrittämistä ei ole tehty Suomen avoimen tiedon käytännöllisestä aspektista.

Myös metatietojen semanttiset ongelmat ovat nousseet vahvemmin esiin nykypäivän ohjelmistojen ja automatiikan kehittyessä (Aroyo *et al.* 2011, s. 65). Avoimen tiedon metatietojen haasteiden ja huomioon otettavien asioiden tunnistaminen on arvokasta *Tiedon hyödyntäminen ja avaaminen* -hankkeelle sekä yleisesti avoimen tiedon hyödyntämiselle. Näiden haasteiden tarkan tunnistamisen puute ja metatietojen haasteiden tutkimuksen vähäisyys tekevät tästä tutkimuksesta arvokkaan.

Tiedon hyödyntäminen ja avaaminen -hankkeen myötä julkisia aineistoja tullaan avaamaan Suomessa yhä enemmän. Avattaville aineistolle luodaan ohjeistuksia ja viitekehyksiä, joita noudattamalla avointa tietoa saadaan hallitusti ja yhdenmukaisesti yhteiskunnalle hyödynnettäväksi (Valtiovarainministeriö 2020). Avoimen tiedon metatiedoissa piilee haasteita, jotka on hyvä saada tiedon avaajille ja hyödyntäjille tietoisuuteen. Näiden haasteiden tunnistamattomuus heikentää avoimen tiedon hyödyntämistä. Lisäksi tämänkaltaisten asioiden tutkiminen lisää avoimen tiedon uskottavuutta, kehitystä ja auttaa perustelemaan tämänhetkistä yhteiskunnallista avoimeen tietoon kohdistuvaa kulttuurinmuutosta. Tämä kokonaisuus toimii tämän tutkimuksen tutkimusongelmana.

Tutkimuskysymyksenä tässä tutkimuksessa on *”Mitä metatietojen haasteita ja huomioon otettavia asioita on olemassa avoimen tiedon avaamisessa ja hyödyntämisessä?”.* Tähän tutkimuskysymykseen pyritään vastaamaan yhdistelemällä teoriaa ja empiiristä tietoa, jota kerätään haastattelujen muodossa. Haastattelut suoritetaan usean eri suomalaisen julkisen organisaation kanssa. Nämä toimijat ovat avanneet julkisesti aineistojaan

ja suhtautuvat positiivisesti tiedon avaamiseen. Näiden organisaatioiden haastateltavat henkilöt antavat uutta tietoa avoimen tiedon haasteista metatietojen näkökulmasta omien kokemustensa sekä oman asiantuntijuutensa kautta.

Diplomityön tavoitteena on edistää hankkeen metatietoja koskevia tavoitetta. Tämä tehdään luomalla uutta tietoa hankkeelle olemassa olevista metatietojen haasteista tietovarantoja avatessa ja hyödynnettäessä. Hankkeen tavoitteita tukeva tieto syntyy tämän tutkimuksen haastattelujen tuloksista ja niistä johdettavista päätelmistä. Tämä tutkimus koostuu rakenteellisesti kuudesta osasta, jotka ovat johdanto, teoreettinen tausta ja lähtökohdat, tutkimusmenetelmä, tulokset ja niiden tarkastelu, pohdinta sekä päätelmät. Tutkimus on laadullinen tapaustutkimus, jossa tiedonkeruumenetelmänä toimii puolistrukturoidut teemahaastattelut.

Johdantokappaleen jälkeen kappaleessa 2 esitetään tämän tutkimuksen lähtölähtökohdat ja teoreettinen tausta. Tutkimuksen kohteena on valtiovarainministeriön *Tiedon hyödyntäminen ja avaaminen* -hanke. Hankkeen taustalla vaikuttaa Suomen tietopoliittiset päätökset sekä Euroopan unionin avoimen datan direktiivi, joka esitellään tarkemmin tämän tutkimuksen lähtökohdissa. Tutkimuksen taustalla vaikuttavien tekijöiden jälkeen käsitellään avoimen tiedon ja metatietojen teoria. Tämän lisäksi tarkastellaan lyhyesti metatietojen laatua sekä metatietojen merkitystä avoimessa tiedossa.

Teoriaosuuden jälkeen käsitellään tarkemmin tutkimuksen menetelmää kappaleessa 3. Tämä tutkimus suoritetaan tapaustutkimuksena, jossa tiedonkeräysmenetelmänä toimii puolistrukturoidut haastattelut. Haastateltavat henkilöt toimivat Suomen julkisissa organisaatioissa, jotka suhtautuvat positiivisesti tiedon avaamiseen. Henkilöillä on kokemusta tiedon avaamisesta, aineistojen metatiedoista tai molemmista edellä mainituista. Haastateltavat organisaatiot ja henkilöt esitetään anonyymisti. Haastattelut analysoidaan induktiivisella analyysimenetelmällä.

Haastatteluista on kerätty tämän tutkimuksen aihepiiriin kuuluvat relevantit löydökset, kuten esimerkiksi avoimen tiedon metatietojen haasteet ja huomioon otettavat asiat. Nämä käsitellään kappaleessa 4. Haastatteluissa ilmenneillä eri löydöksillä on selkeät syy-seuraussuhteet. Nämä havainnollistetaan myöhemmin kuvassa 1.

Haastattelujen tulosten tarkastelun jälkeen kappaleessa 5 pohditaan tuloksia. Kappaleessa peilataan saatuja tuloksia aiempiin tutkimuksiin sekä aiemmin käsiteltyyn teoriaan. Tämän jälkeen kappaleessa 6 esitetään tämän tutkimuksen päätelmät ja vastataan tutkimuskysymykseen. Lopuksi esitetään lyhyesti tutkimuksen arviointi ja mahdolliset jatkotutkimusaiheet.

2. TEOREETTINEN TAUSTA JA LÄHTÖKOHDAT

Tämän tutkimuksen lähtökohtana toimii valtiovarainministeriön aloittama *Tiedon hyödyntäminen ja avaaminen* -hanke. Tässä luvussa tutustutaan hankkeen periaatteisiin ja Euroopan unionin avoimen tiedon direktiiviin. Lisäksi käsitellään avointa tietoa ja metatietoja koskeva teoriapohja.

2.1 Tiedon hyödyntäminen ja avaaminen -hanke

Valtiovarainministeriön aloittaman T*iedon hyödyntäminen ja avaaminen* -hankkeen tarkoituksena on kehittää Suomen julkisten ja julkisomisteisten organisaatioiden tiedon hyödynnettävyyttä ja avoimuutta. Hankkeen perimmäisenä tavoitteena on tukea Suomen kansallisen tietopolitiikan kantavia periaatteita. Lisäksi hankkeella tuetaan julkista hallintoa sekä avoimen datan direktiivin (PSI-direktiivi) toimeenpanoa. Hankkeen toimikausi on vuoden 2020 alusta vuoden 2022 loppuun, jonka aikana edistetään julkisen tiedon entistä laajempaa ja tehokkaampaa hyödyntämistä koko yhteiskunnassa. (Valtiovarainministeriö 2020)

Tiedon avaaminen ja hyödyntäminen vaatii sitovien laatukriteerien laatimista ja suunnittelua, sekä julkisten tietojärjestelmien lähdekoodin ensisijaistamista. Lisäksi vaaditaan laaja verkko yhteistyötahoja luomaan vahvoja toimintaehdotuksia, uutta tietoa ja valmistelemaan toimenpiteitä hankkeen tavoitteiden saavuttamiseksi. Hankkeessa ovat mukana muun muassa Digi- ja väestötietovirasto, Tilastokeskus ja Avoindata.fi. Muita keskeisiä toimijoita ovat valtionhallinto, eri kunnat ja kuntayhtymät sekä julkishallinnon omistamat yritykset. (Valtiovarainministeriö 2020)

Hanke on jaettu neljään eri työpakettiin, jotka kaikki pureutuvat erilaisiin ja eri tasoisiin hankkeen tavoitteisiin ja haasteisiin. Työpaketti 1 *Strategiset tavoitteet* pyrkii syventämään ja luomaan toimintaperiaatteita kansalliselle tietopolitiikalle. Työpaketti 2 *Tiedon saatavuus* tunnistaa hyötypotentiaalisia tietoja ja luo mallin niiden avaamiseen ja julkaisemiseen. Työpaketti 3 *Tiedon laatu* määrittelee ja käyttöönottaa tiedon laatukriteerit, jotka tulevat käyttöön kansallisesti. Viimeinen työpaketti 4 *Tiedon semanttinen ja tekninen yhteentoimivuus* laatii API-linjaukset tiedon jakamiselle sekä luo mallin luodun yhteentoimivuusalustan tiedon hyödyntämiselle ja avaamiselle. (Valtiovarainministeriö 2020) Tämä diplomityö tukee ja tuo uutta tietoa kaikkien työpakettien käyttöä varten metatietojen haasteista ja huomioon otettavista asioista.

Hankkeen yhtenä tavoitteena on parantaa hyödynnettävän ja avattavan tiedon laatua sekä yhteentoimivuutta teknisesti ja semanttisesti. Hanke tukee tiedon löydettävyyden ja ymmärrettävyyden parantamista kehittämällä erityisesti tietoaineistojen metatietoja ja rajapintadokumentaatiota. Hanke tulee tukemaan tietojen semanttista yhteentoimivuutta edistämällä tietosisältöjen ja metatietojen yhtenäistämistyötä. Semanttinen yhteentoimivuus tukee tiedon merkityksen säilymistä, ymmärrettävyyttä, hyödynnettävyyttä ja löydettävyyttä. (Valtiovarainministeriö 2020)

Tällä hetkellä merkittävä osa avoimen tiedon tietovarantojen metatiedoista kirjoitetaan käsin. Tämä aiheuttaa ongelmallisia vapausasteita käytettyihin metatietoihin, mikä johtaa pahimmillaan heikkoon tiedon laatuun, jonka vuoksi tiedot eivät ole helposti hyödynnettävissä.

Tämän diplomityön näkökulmasta tärkein hankkeen osatavoite on tunnistaa avoimen tiedon metatiedon kehittämistarpeet, jotta tiedon semanttinen ja tekninen yhteentoimivuus paranee. Tätä tavoitetta varten avoimen tiedon metatietoihin kohdistuvat muutostarpeet on tunnistettava ja dokumentoitava (Valtiovarainministeriö 2020).

2.2 Euroopan unionin avoimen datan direktiivi

Euroopan unionin avoimen datan direktiivi (nk. PSI-direktiivi) luotiin parantamaan Euroopan informaatiotarjontaa. Se kannustaa Euroopan valtioita kehittämään avoimeen tietoon liittyvää politiikkaa ja säännöksiä ja sen tavoitteena on asettaa julkisen sektorin tietoja paremmin koko yhteiskunnan hyötykäyttöön. Toinen tärkeä tavoite on pyrkiä parantamaan avoimen tiedon uudelleenkäyttöä. Direktiivin ansiosta avoimen tiedon uudelleenkäyttäjien on helpompi löytää arvokkaita tietoaineistoja (Janssen 2011). Avoimen tiedon aineistojen löydettävyyttä pyritään myös parantamaan *Tiedon avaamisen ja hyödyntämisen* -hankkeen kautta.

Tuore uudelleenlaadittu PSI-direktiivi (EU 2019/1024) pyrkii edistämään julkisen sektorin tuottaman tiedon täysmääräistä hyödyntämistä. Direktiiviä on laajennettu koskemaan kattavammin kaikkien jäsenvaltioiden julkisen sektorin hallussa olevien tietojen säädöksiä. Uudet direktiivin muutokset keskittyvät pääosin reaaliaikaisen datan tarjoamiseen asianmukaisilla menetelmillä, arvokkaan julkisen tiedon tarjonnan lisäämiseen sekä vanhan PSI-direktiivin täsmentämiseen. Direktiivi antaa jäsenvaltioiden hoitaa käytännön järjestelyt säädöksien toteuttamiseksi esimerkiksi metatietokatalogien suhteen. (Forsström 2019)

Uudistettu PSI-direktiivi nostaa esiin kuusi tietovarantotyyppiä, joita pidetään erityisen arvokkaina avointa dataa ajatellen. Näiden tietovarantotyyppien tietojen avaamisella ja

tietojen uudelleenkäytöllä on tärkeitä potentiaalisia hyötyjä taloudelle ja yhteiskunnalle, sillä tietojen avaaminen ilmaiseksi koneluettavassa muodossa asianmukaisen ohjelmointirajapinnan (API) kautta luo valtavasti uusia mahdollisuuksia tiedon täysimääräiseen hyödyntämiseen.PSI-direktiivin tietovarantotyypit ovat seuraavat (European Comission 2020):

1. Paikkatiedot
2. Maan havainnointi ja ympäristö
3. Säätiedot
4. Tilastotieto
5. Yritys- ja yritysten omistustiedot
6. Liikkuvuustiedot

Tässä tutkimuksessa pyritään keskittymään juuri näiden tietovarantotyyppien hyödyntämiseen ja avaamiseen. Tietovarantotyypit ovat keskeisessä asiassa myös silloin, kun selvitetään haastattelujen kautta metatietojen haasteita.

2.3 Avoin tieto

Avoimella tiedolla tarkoitetaan eri tietoaineistossa olevaa tietoa, joka on tarkoitettu vapaasti kaikille jaettavaksi ja hyödynnettäväksi ilmaiseksi mihin tahansa käyttötarkoitukseen. Avoimen tiedon tietotyyppi voi olla mitä vain koneellisesti luettavaa tietoa, kuten esimerkiksi kuvia, tekstiä tai tilastollista tietoa. Tarkkaa yksiselitteistä määrittelyä avoimelle tiedolle ei ole olemassa, mutta useimmiten termi käsitetään hyvin samankaltaisesti ympäri maailmaa. Termi *avoin tieto* on hyvin lähellä termiä *avoin data*. Puhekielessä termeillä tarkoitetaan yleisesti samaa asiaa, ja tästä johtuen termit sekoittuvat usein keskenään eri lähteissä. Suomessa YSO – Yleinen suomalainen ontologia käyttää termiä *avoin tieto*, ja tätä käytetään myös tässä tutkimuksessa (Finto 2020)

Avoin tieto ajatuksena lähti liikkeelle 1980-luvulla vapaita ohjelmistoja ja avoimia lähdekoodeja edistävien liikkeiden kautta. Myöhemmin avoin tieto levisi laajemmin yleiseen tietoisuuteen maailmassa, ja on sittemmin noussut globaaliksi supertrendiksi. Suomessa tiedon avaaminen alkoi kiihtyä vuonna 2009, kun Euroopan unionin PSI-direktiivi julkais-

tiin. Kiihtyminen jatkui edelleen vuonna 2011, kun Suomen hallitus julkisti periaatepäätöksen, jonka seurauksena Suomen tietoaineistojen avoimuutta, uudelleenkäytettävyyttä ja maksuttomuutta korostettiin. (Avoindata.fi 2021)

Avoimessa tiedossa avattavaa tietoa kutsutaan *aineistoiksi* tai *kokoelmiksi*. Kokoelmat voivat pitää sisällään useita eri aineistoja, mutta niitä voidaan itsessään kutsua myös aineistoksi. Avoimessa tiedossa termillä *lisenssi* viitataan lailliseen lisenssiin, jolla aineisto on julkaistu. Erillisen lisenssin puuttuessa, tulkitaan tämän viittaavaan vallitsevaan lainsäädäntöön, jonka mukaisesti aineisto on julkaistu (Open Knowledge Fountain 2021). Tässä tutkimuksessa kaikesta avattavasta tiedosta käytetään termiä *aineisto*. Aineisto voi olla muutakin kuin avattavaa taulukkodataa. Se pitää sisällään kaiken avattavan tiedon, kuten esimerkiksi erinäisiä kuvauksia, julkaisuja, sanastoja sekä luokittelutietoja tai muita hyödynnettävissä olevia tietoja. On hyvä huomata, että *tietovarantotyypillä* ja *aineistotyypillä* tarkoitetaan eri asioita. Aineistotyypillä voidaan viitata avattavan tiedon aineiston eri muotoihin, joita mainittiin edellä. Tietovarantotyypeillä puolestaan kuvataan 2.2 kappaleessa olevien kuuden eri kategorian tavoin avoimen tiedon aineiston sisältämän tiedon tyyppiä.

Yleisesti avoimeksi tiedoksi voidaan tulkita sellainen tieto, joka täyttää seuraavat kahdeksan periaatetta (Tauberer 2014a):

1. *Täydellisyys* – Kaikki sellainen julkinen tieto, joka ei ole salattua, on avattu kokonaisuudessaan.
2. *Ensisijaisuus* – Tieto on avattu mahdollisimman läheltä sen lähdettä ilman liiallisia muunnoksia tai yhteenkokoamista.
3. *Ajankohtaisuus* – Tieto on avattu tarpeeksi nopeasti ylläpitääkseen avattavan tiedon arvon.
4. *Käytettävyys* – Tieto on mahdollisimman laajalle yleisölle ja käytölle avattuna.
5. *Koneluettavuus* – Tieto on koottu ja avattu sellaisessa muodossa, että se on koneluettavissa automatisoituja prosesseja varten.
6. *Ei-syrjivä* – Tieto on kaikkien käytettävissä anonyymisti.
7. *Ei-poissulkeva* – Tieto on julkaistu niin, että kaikilla on yhtä hyvät mahdollisuudet käyttää tietoa. Yksittäisillä tahoilla ei saa olla ylilyöntiasemaa tiedon käyttämisessä.

8. *Lisenssivapaa* – Tieto ei saa olla minkään tekijänoikeusmerkin, patentin tai salassapitosopimuksen alaisena.

Muita merkittäviä periaatteita ovat (Tauberer 2014b):

- *Maksuttomuus verkossa* – Tiedon voidaan katsoa olevan avointa, kun se on helposti käytettävissä ja löydettävissä verkossa.
- *Pysyvyys* – Tieto pitää olla pysyvästi ja pitkäaikaisesti jaettuna samassa muodossa ja paikassa.
- *Luotettavuus* – Tieto on läpinäkyvästi julkaistu, eikä sitä voida muokata jälkikäteen.
- *Avoimuuden oletettavuus* – Tiedon oletetaan olevan avoimena alusta lähtien. Sille tehdään proaktiivisia toimintoja, jotta tieto on vapaasti käytettävissä mahdollisimman esteettömästi ja nopeasti verkossa.
- *Dokumentaatio* – Tieto on dokumentoitu tarpeeksi kattavasti niin, että se on mahdollisimman hyvin hyödynnettävissä.
- *Turvallisesti avattava* – Tieto on avoimena niin, että se on aina turvallisesti avattavissa ilman tietoturvallisuuden riskejä.
- *Suunniteltu yleisölle* – Tieto avataan ymmärrettävässä muodossa, jotta se tuottaa mahdollisimman paljon arvoa yhteisölle.

Suomessa avointa tietoa on saatavilla yhä enenevissä määrin. Yleisesti ottaen avoin tieto noudattaa edellä olevia periaatteita. Tietyissä tapauksissa kuitenkin eri aineistojen saatavuutta on rajoitettu GDPR- tai turvallisuussyistä. Suomessa avoimet aineistot noudattavat lähtökohtaisesti *JHS 189 Avoimen tietoaineiston käyttölupa* -suositusta, jonka mukaan avattuja aineistoja saa vapaasti käyttää kaikin mahdollisin tavoin, kunhan aineiston alkuperäinen lähde on mainittuna (Suomidigi 2020).

Avoimen tiedon arvo tulee esiin tietoa käyttäessä. Ennen tiedon käyttöä se on jaettava ja löydettävä. Aineiston voidaan katsoa olevan avointa, jos sen jakelutavat täyttävät seuraavat 11 ehtoa (Open Knowledge Fountain 2021):

1. *Saavutettavuus* - Aineiston pitää olla kokonaisuudessaan saatavilla enintään kohtuullisella luovutuskustannuksella, ja mieluiten maksutta ladattavissa internetin kautta. Aineiston pitää myös olla saatavilla käytännöllisessä ja muokattavassa muodossa.
2. *Uudelleenjakelu* - Lisenssi ei saa rajoittaa ketään myymästä tai antamasta aineistoa yksinään tai osana kokoelmaa, joka sisältää aineistoja useista eri lähteistä. Lisenssin mukaisesti ei voida vaatia rojalteja tai muita myyntiin tai jakeluun kohdistuvia maksuja.
3. *Uudelleenkäyttö* - Lisenssin on sallittava muokkaus ja muokattujen aineistojen jakelu alkuperäisen aineiston ehdoilla. Lisenssi voi sisältää vaatimuksia aineiston eheydestä ja viittaamisesta alkuperäiseen aineistoon: katso alla periaate 5 (*Viittaaminen*) ja periaate 6 (*Eheys*).
4. *Vapaa teknisistä rajoitteista* - Aineisto pitää tarjota sellaisessa muodossa, ettei yllä mainittujen kohtien mukaiselle toiminnalle ole teknisiä esteitä. Tämä voidaan saavuttaa tarjoamalla aineisto avoimessa formaatissa, kuten sellaisessa, jonka spesifikaatio on julkisesti ja vapaasti saatavilla, ja joka ei aseta rahallisia tai muita rajoitteita formaatin käytölle.
5. *Viittaaminen* - Lisenssi voi vaatia uudelleenjakelun ja uudelleenkäytön ehtona, että aineiston tekijät mainitaan. Jos tällainen ehto asetetaan, sen noudattaminen ei saa aiheuttaa kohtuuttomasti työtä. Jos viittaamista uudelleenkäytettävän aineiston edelliseen tekijään vaaditaan, tulee aineiston yhteydessä toimittaa lista niistä, jotka pitää mainita aineiston tekijöinä.
6. *Eheys* - Lisenssi voi vaatia muokatun aineiston jakelun ehtona, että uusi aineisto nimetään eroavasti tai uudella versionumerolla alkuperäiseen aineistoon nähden.
7. *Ei henkilöiden tai ryhmien syrjintää* - Lisenssi ei saa asettaa henkilöitä tai ryhmiä eriarvoiseen asemaan.
8. *Ei syrjintää käyttökohteiden suhteen* - Lisenssi ei saa rajoittaa ketään käyttämästä aineistoa jollakin määrätyllä käyttöalueella. Se ei saa esimerkiksi estää aineiston kaupallista käyttöä tai käyttöä geenitutkimukseen.
9. *Lisenssin jakelu* - Aineistoon liittyvien oikeuksien tulee koskea kaikkia, joille aineisto on jaeltu ilman tarvetta erillisten lisenssien käyttöön.
10. *Lisenssi ei saa olla kokoelmakohtainen* - Aineistoon liittyvät oikeudet eivät saa olla riippuvaisia tiettyyn kokoelmaan kuulumisesta. Jos aineisto irrotetaan tästä

kokoelmasta ja sitä käytetään tai jaellaan aineiston lisenssin mukaisesti, niin kaikilla osapuolilla, joille se on uudelleenjaeltu, tulee olla samat oikeudet, jotka myönnettiin alkuperäisen kokoelman yhteydessä.

11. *Lisenssi ei saa rajoittaa muiden aineistojen jakelua* - Lisenssi ei saa asettaa rajoituksia muihin aineistoihin, joita jaellaan yhdessä lisensoidun aineiston kanssa. Lisenssi ei saa esimerkiksi vaatia, että kaikki muut sen yhteydessä jaettavat aineistot olisivat avoimia.

Linkitetty tai yhdistetty avoin tieto (engl. linked open data) korostaa avoimen tiedon edellä mainittujen ominaisuuksien lisäksi tiedon yhdistettävyyttä ulkopuolisiin aineistoihin verkossa. *Yhdistetty avoin tieto* voi terminä tarkoittaa kokoelmaa yleisistä käytännöistä jäsenneltyjen tietojen linkittämisestä toisiinsa verkossa. (Attard et al. 2016)

Yhdistetyn avoimen tiedon potentiaali on valtava, ja se on ensimmäinen askel kohti semanttista webiä (engl. Web of Data, Semantic Web). *Semanttisella webillä* tarkoitetaan tavoitetilaa, jossa verkossa olevat aineistot ovat yhdistettynä toisiinsa luoden valtavan yleisen hajautetun kokoelman aineistoista, joka korvaisi yksittäiset toisistaan eristetyt tietolähteet, joita tällä hetkellä yleisesti käytetään. (Attard et al. 2016)

Yhdistetty avoin tieto on vaativa kehityskohde, jota pyritään saavuttamaan globaalisti. Konkreettisia hyötyjä syntyy merkittävä määrä, mikäli yhdistetyn avoimen tiedon käyttö saadaan saavutettua (Attard et al. 2016):

1. Yhdistetyn avoimen tiedon luoma yksinkertaistettu pääsy tiedon hyödyntämiseen yhtenäisen tietomallin avulla helpottaa tiedon käyttöä valtavasti.
2. Aineistojen sisällön kattava esitys mahdollistaa tehokkaan semanttisen dokumentoinnin, joka auttaa aineiston ymmärrettävyyttä.
3. Olemassa olevien sanastojen uudelleenkäyttäminen.
4. URI:en kohdennettu käyttäminen auttaa tietojen yhdistettävyydessä ja viittauksissa. Tämä luo läpinäkyvyyttä tiedon käyttämiselle ja helpottaa tiedon ymmärrettävyyttä.
5. Tietojen yhdistettävyys ja niiden linkit mahdollistavat aineiston asiaan liittyvien tietojen yhtenäisen hyödyntämisen.

Näiden hyötyjen saavuttaminen nostaa avoimen tiedon arvoa ja hyödynnettävyyttä entisestään.

2.4 Metatieto

Metatieto on yksinkertaistettuna tietoa tiedosta. Sillä on useita eri määrittelyjä, jotka muuttuvat kontekstin mukaisesti, mutta yleisesti ottaen sen voidaan ajatella olevan kuvaus jostakin informatiivisesta tietosisällöstä (Pomerantz 2015). Metatiedon tarkoitus on tuottaa riittävää ja oikeellista informaatiota käyttäjälle, jotta käyttäjälle voidaan antaa todenmukainen kuva kyseisestä tietosisällöstä ilman sen avaamista (Greenberg et al. 2008). Laadukas metatieto antaa tarkkaa ja kuvaavaa tietoa kyseessä olevasta tietosisällöstä, joka edesauttaa tiedon hyödynnettävyyttä, löydettävyyttä ja jatkokäyttöä.

Metatieto on vahvasti mukana nykyisissä tietojärjestelmissä. Valtaosa ohjelmistoista on metatietovetoisia. Tämä parantaa tietojärjestelmien ominaisuuksia ja toimivuutta kokonaisvaltaisesti esimerkiksi tietoa etsiessä, tallentaessa tai jakaessa (Riley 2017). Metatieto toimii tiedon taustalla, joka auttaa tiedon tulkitsemista koneellisesti sekä ihmisen toimesta.

Metatiedot voidaan jakaa eri kategorioihin usealla eri tavalla. Yksi tapa on jakaa metatiedot seuraaviin kategorioihin: kuvaileva-, hallinnollinen-, rakenteellinen metatieto sekä metatietojen merkintäkielet. *Kuvaileva metatieto* kuvaa sananmukaisesti kohteena olevaa tietoa, mikä edesauttaa löytämään ja ymmärtämään kyseistä tietoa. *Hallinnollinen metatieto* puolestaan helpottaa tiedon käyttämistä, tallentamista ja säilömistä (Gilliland 2008). Nämä tiedot sisältävät omistajuuteen, oikeuksiin ja teknisiin ominaisuuksiin liittyviä metatietoja. *Rakenteellinen metatieto* antaa puolestaan tietoa tietohierarkioista ja yhteyksistä muihin tietoihin, joka auttaa etenkin eri tietovarantojen välisessä navigoinnissa. Teorian mukaan m*etatietojen merkintäkielillä* erotetaan metatietojen rakenteellinen osuus sen semanttisesta ja informatiivisesta sisällöstä. Merkintäkielillä siis kuvataan metatietoja teknisesti ilman kuvailua ja ihmisen helposti ymmärrettävää kieltä. Merkintäkieliä on olemassa useita, joista käytetyin on XML (engl. eXtensible Markup Language), joka julkaistiin vuonna 2000 (Riley 2017).

Edellä olevat neljä metatiedon kategoriaa voidaan havainnollistaa käytännön esimerkkinä. Tässä kohtaa kuvitellaan aineiston olevan albumi. Samoin kuin aineistot, albumit voivat sisältää kaikkia neljää eri metatietotyyppiä. Kuvaileva metatieto kuvailee albumin perustietoja, jotka helpottavat albumin tunnistuksessa, ja auttavat ymmärtämään mitä albumi sisältää. Perustietoja ovat esimerkiksi albumin nimi, artisti ja julkaisija. Hallinnollinen metatieto sisältää puolestaan albumin tekijänoikeus- tai omistajuustietoja sekä tietoa albumin jakamisoikeuksista. Rakenteellinen metatieto kuvaa puolestaan albumin tietohierarkiaalista järjestystä, kuten esimerkiksi tietoa siitä, mihin tuoteperheeseen albumi kategorisesti kuuluu. Tämä tieto helpottaa albumin löydettävyyttä ja järjestettävyyttä,

sekä antaa kontekstia albumin käyttötarkoituksesta. Albumin metatiedon merkintäkielen voidaan osaltaan ajatella pitävän sisällään taulukkomaisesti kaikki tärkeät metatiedot, jotka ovat teknisesti luettavissa. Tämä helpottaa albumin metatiedon käyttämistä, hyödyntämistä ja yhteentoimivuutta.

Seuraavaksi todettakoon, että metatieto koostuu rakenteellisesti sen sisältämistä skeemoista, elementeistä ja arvoista. *Skeemat* kuvaavat metatiedon rakennetta, *elementit* sen sisältämiä osia ja *arvot* metatiedon sisältämää tietosisältöä (Pomerantz 2015). Metatiedon skeemaa voidaan kutsua myös metatiedon standardiksi, joka määrittelee mitä elementtejä metatieto sisältää. Elementit voivat yksinkertaistetusti olla metatiedon eri arvoparametrien nimiä. (Zhang & Gourley 2008). Käyttäen edellistä havainnollistavaa esimerkkiä, musiikkialbumin metatietoskeema voisi kuvitteellisesti käsittää sen takakannesta löytyvät perustiedot. Skeeman elementit voisivat tällöin olla esimerkiksi artistitieto, albumin nimi sekä listaus kappaleista. Arvot olisivat tällöin puolestaan artistin, albumin sekä kappaleiden nimet.

Metatieto on hyvin vaivalloista ja kallista kehittää sekä ylläpitää. Kuitenkin nykymaailmassa internetin seurauksena, on sen merkitys kasvanut merkittävästi. Suurimpia metatiedon hyötyjä on sen vaikutus tiedon löydettävyyteen. Se mahdollistaa myös tiedon yhteneväisyyksiä eri kokoelmien läpi, jos kuvaileva metatieto on yhdistettävissä eri aineistoissa (Gilliland 2008). Kuvailevan metatiedon skeemoilla, eli standardeilla ja semanttisuudella, on suuri merkitys aineiston löydettävyyteen.

Löydettävyyden lisäksi metatiedoilla on myös suuri rooli, kun luodaan ja ylläpidetään aineistojen kontekstia sekä suhteita ympäröiviin tekijöihin (Gilliland 2008). Ilman metatietoa aineistot olisivat vaikeasti yhdistettävissä mihinkään käyttötarkoitukseen tai lähteeseen, jolloin niiden luotettavuus olisi heikkoa sekä mahdollinen käyttötarkoitus epäselvä.

Löydettävyyden ja yhdistettävyyden lisäksi metatiedoilla on suuri rooli pitää aineistot eheänä eri migraatioiden läpi tietoa uudistettaessa tai siirrettäessä ohjelmistojen kautta. Tekninen ja kuvaileva metatieto muistaa mitä aineiston tieto on, mistä se on tullut ja mihin se liittyy. On kuitenkin hyvä ylläpitää itse aineiston lisäksi sen sisältämää metatietoa, sillä ilman sitä aineisto voi irtaantua sitä kuvaavista metatiedoista. Näin aineiston konteksti, sen sisältämä tieto sekä käyttötarkoitus häviää (Gilliland 2008).

National Information Standards Organization (NISO) on voittoa tavoittelematon valtuutettu organisaatio, joka on perustettu 1939. NISO ylläpitää, tutkii ja julkaisee yleiskäytännöllisiä julkisia tiedonkäytön standardeja aktiivisesti yhdessä muiden informaatiotekniikan yhteisöjen kanssa (National Information Standards Organization 2021, s. 63). NISO

julkaisi 2007 viitekehyksen hyvien aineistojen luomiselle, jossa lueteltiin kuusi hyvän metatiedon periaatetta:

1. Metatieto on yhdenmukainen olemassa olevan ja yleisesti käytössä olevien standardien kanssa. Metatiedon standardi on valittu sopimaan sen tietosisällön kanssa ja on yhteentoimiva aineiston käyttämisen kanssa nyt ja tulevaisuudessa. Metatiedon standardi ottaa huomioon myös käyttäjät.
2. Metatieto tukee yhteentoimivuutta. Aineistoja on vaikea käyttää suoraan, jos niiden metatieto on laadittu liian paikallisesta näkökulmasta. Laajemmassa käytössä metatiedon olisi hyvä olla mahdollisimman yleispätevää globaalisti.
3. Metatiedossa on käytössä kontrolloituja standardeja, sanastoja sekä strukturoituja skeemoja aineiston sisällön kuvailuun. Käytettyjen termien ja sanastojen käyttäminen on tarkkaan harkittu ja dokumentoitu.
4. Metatieto sisältää selkeät käyttöehdot sekä oikeus- ja vastuulausekkeet aineiston käyttämisestä ja sen rajoituksista.
5. Metatieto on laadittu tukemaa aineiston pitkän aikavälin hallintaa, säilömistä ja sen käytön tukemista. Hallinnoiva metatieto on tarkoitettu kuvaamaan miten ja koska aineisto on luotu ja kuka on siitä vastuussa.
6. Metatieto on osa aineistoa. Hyvän ja laadukkaan aineiston olemassaolon ehtona on hyvä ja laadukas metatieto. Metatieto vaatii itsestään myös tietoa ja kuvailua, niin kutsuttua meta-metatietoa, joka kuvailee mitä metatieto pitää sisällään.

Näiden periaatteiden noudattamisella voidaan avata tietoa yleisellä tasolla niin, että avoin tieto on yhteentoimiva globaalisti.

2.5 Metatiedon laatu

Laadukas metatieto on täydellistä, oikeellista ja relevanttia. Täydellinen metatieto kuvaa kokonaisvaltaisesti ja riittävällä tarkkuudella kyseessä olevan metatiedon kaikkia kenttiä ja sisältöä. Se käsittää myös metatietojen puutteet ja kattaa sen, mitä tietoa tietueessa on oltava, jotta sen voi käsittää olevan täydellistä (Greenberg et al. 2008). Metatiedot sisältävät usein joukon pakollisia ja vapaaehtoisia kenttiä, joiden täydellisyys vaikuttaa suoraan metatiedon laatuun.

Metatiedon oikeellisuudella tarkoitetaan sitä, pitävätkö metatiedon esittämät asiat paikkansa. Tämä voidaan jakaa kahteen tasoon, joista ensimmäinen alempi taso käsittelee tiedon oikeellisuutta kieliopillisesti ja syntaktisesti. Oikeellisuuden korkeampi toinen taso

käsittelee metatiedon semanttista oikeellisuutta. Kieliopillisesti ja syntaktisesti laadukkaan metatiedon kaikki osat sisältävät luettavaa standardimaista tietoa, joka ei pidä sisällään virheellisiä asioita kuten kirjoitusvirheitä, vääriä formaatteja tai merkkejä. Semanttisesti oikeellinen metatieto on sellaista, jossa tieto on esitetty oikein niin, että se kuvastaa todellisuutta selkeästi, eikä sitä voi helposti tulkita virheellisesti. (Greenberg et al. 2008) Semanttinen oikeellisuus on haastavaa varmistaa ja todentaa, sillä tiedon semanttisuus on usein subjektiivinen asia, toisin kuin syntaksinen oikeellisuus, joka voidaan varmistaa helpommin.

Metatiedon semanttisuus tuo haasteita metatiedon käsiteltävyyteen, sillä metatietoja hyödynnetään nykyään pääasiassa ohjelmallisesti. Metatietoa on kuitenkin muussakin muodossa kuin digitaalisessa. Tietoa tiedosta voi syntyä ja siirtyä esimerkiksi fyysisten kirjoitusten tai ihmisten puheessa (Gilliland 2008).

Metatiedon relevanttius on vahvasti riippuvainen tiedon kontekstista. Metatiedon laatu voi olla täydellistä ja oikeellista, mutta se voi olla käyttökelvotonta puutteellisen relevanttiuden seurauksena. Käytännössä tämä tarkoittaa sitä, että metatiedon kentät voivat olla oikeellisesti täytetty ja kokonaisuus voi olla täydellinen, mutta laatu voi olla silti heikkoa, jos metatiedot eivät ole yhteydessä kyseessä olevaan tietoon. Yksi tapa välttää epärelevantin metatiedon syntymistä on rajoittaa kenttien täyttämistä. Näin subjektiiviset relevanttiuden ongelmat voidaan siirtää oikeellisuuden puolelle. (Greenberg et al. 2008)

2.6 Metatieto avoimessa tiedossa

Metatiedon arvo korostuu entisestään, kun puhutaan avoimesta tiedosta. Etenkin julkisten organisaatioiden avatun tiedon metatieto on kriittistä, sillä se on avainasemassa aineiston läpinäkyvyyden ja luotettavuuden todistamisessa. Sen tehtävä on myös antaa oikea käsitys metatiedoista tutkijoille ja muille käyttäjille (Aroyo *et al.* 2011, 66).

Avoimen tiedon metatietoa ja sen laatua on tutkittu julkisten organisaatioiden avoimien portaalien näkökulmasta. Avoimen tiedon metatietoa on vaikea arvioida ja tutkia sen multidimensionaalisuuden ja erilaisten käyttäjien seurauksena. Julkiset organisaatiot ovat havainneet avoimen tiedon aineistojen ja metatietojen ongelmia, ja niihin on pyritty löytämään ratkaisuja. (Kubler *et al.* 2018)

Metatiedon puuttuminen hankaloittaa suoraan aineiston löydettävyyttä (Neumaier et al. 2016). Lisäksi metatiedoilla on selvä yhteys aineistojen ymmärrettävyyteen ja tietouden lisäämiseen siitä, miten aineisto on luotu (Sugimoto 2014). Käytännössä metatietojen käyttö on kuitenkin haastavaa yleispätevien metatietosäännösten puuttumisen takia (Zuiderwijk *et al.* 2012).

Avoimen tiedon metatiedon on hyvä olla kytköksissä internetin välityksellä muihin yleisiin palveluihin. Avoimelle tiedolle on olemassa useita viitekehyksiä kuvaamaan metatietoa ja aineistoa. Yksi näistä on World Wide Web Consortiumin (W3C) standardoima RDF (engl. Resource Description Framework), joka on rakennettu nimenomaisesti web-ympäristössä tapahtuvaan tiedonsiirtoon (W3C 2014).

Eri viitekehyksiä käyttäessä ja yleisestikin avointa tietoa tarkastellessa on hyvä käyttää yleisiä sanastoja, jotta voidaan parantaa eri aineistojen ja palveluiden semanttista yhteentoimivuutta. *Sanastolla* tarkoitetaan ohjeistusta käyttää tiettyjä termejä ja ilmauksia aineiston kuvailuun. Se voi määritellä esimerkiksi metatietojen elementtien nimiä. Yksi yleisimmistä sanastoista on W3C:n vuonna 2020 kehittämä DCAT-2 (W3C 2020). DCAT on avoimen tiedon julkaisua auttava yleinen sanasto, joka auttaa tiedon löydettävyyttä ja ymmärrettävyyttä. Lisäksi se tarjoaa kuvauksen aineistosta sekä esimerkkejä sen käytöstä.

Avoin tieto on yleensä käytettävissä erilaisten portaalien kautta. Esimerkiksi Suomessa yksi tunnetuimmista avoimen tiedon portaaleista on avoindata.fi. Avoimen tiedon portaalien metatietojen laatua voidaan arvioida seuraavan viiden kriteerin avulla, jotka on rakennettu DCAT-sanastoa hyödyntävän aineiston sisällön arvioimisen pohjaksi (Neumaier et al. 2016, s. 10):

1. *Olemassaolo* – Sisältääkö metatieto kaikki tarvittavat kentät?
2. *Vaatimustenmukaisuus* – Käyttääkö ja noudattaako metatieto annettuja vaatimuksia ja standardeja?
3. *Saatavuus* – Onko metatiedot käytettävissä ja saatavissa?
4. *Tarkkuus* – Kuvaako metatiedon tieto aineistoa riittävällä tasolla?
5. *Avoin tieto* – Onko metatieto laadittu niin, että se on sopiva avoimelle tiedolle?

Tämän kaltaisten viitekehysten ja sanastojen kriteerien noudattamisen avulla voidaan saavuttaa tilanne, jossa avoimen tiedon aineisto on hyvin löydettävissä ja hyödynnettävissä.

3. TUTKIMUSMENETELMÄ

Tässä diplomityössä kerätään uutta tietoa avoimen tiedon metatietojen haasteista ja huomioon otettavista asioista, sekä tehdään päätelmiä yhdistelemällä kerättyä tietoa ja teoriaa. Tämä tutkimus on laadullinen tapaustutkimus, jossa tiedonkeruumenetelmänä toimii puolistrukturoidut haastattelut. Haastattelujen kautta laadullista tietoa kerätään haastateltavien henkilöiden asiantuntijuuden ja kokemusten pohjalta tapauskohtaisesti. Tarkoituksena on kerätä tietoutta avoimesta tiedosta mahdollisimman kattavasti eri näkökulmista.

Hyvin usea laadullinen tutkimus on tapaustutkimus (case study). Tapaustutkimuksessa pyritään luomaan ymmärrystä jostakin laajemmasta ilmiöstä tai asiasta. Tutkittavana tapauksena voi olla jokin organisaatio, prosessi tai tapahtuma (Kallinen & Kinnunen 2021). Tässä tutkimuksessa tapauksina nähdään haastateltavien henkilöiden käsitys vallitsevasta avoimen tiedon metatietojen tilanteesta.

Haastateltavina kohteina ovat eri julkiset ja julkisomisteiset organisaatiot, jotka suhtautuvat positiivisesti avoimen tiedon jakamiseen. Kustakin kohdeorganisaatioista pyritään valitsemaan yhdestä kolmeen haastateltavaa, jotka ovat olleet tekemisissä avoimen tiedon jakamisen ja metatietojen parissa. Kohdeorganisaatiot pyritään valitsemaan niin, että ne edustavat erilaisia organisaatioita, tai joiden jakamat aineistot ovat toisistaan mahdollisimman erilaisia. Näin kerätyt tiedot eri tapauksista tulevat laajalta kirjolta ja tutkimuksen validiteetti paranee.

Puolistrukturoidussa haastattelussa, eli teemahaastattelussa, tietoa kerätään valitusta aihepiiristä. Toisin kuin strukturoidussa haastattelussa, teemahaastattelut suoritetaan ilman yhdenmukaista valmiiksi määrättyä kysymyslistaa. Teemahaastattelu sopii hyvin laadullisen tutkimuksen tiedonkeräämiseen. Valitun aihepiirin tiimoilta voidaan kysyä suuntaa antavia ennalta määriteltyjä kysymyksiä vapaassa järjestyksessä. Teemahaastattelun aikana on myös mahdollista kysyä tarkentavia kysymyksiä annettuihin vastauksiin (Saunders et al. 2019, s. 437). Tässä tutkimuksessa annettiin puolistrukturoitu alustava kysymyslista etukäteen kaikille haastateltaville henkilöille, joka on nähtävissä liitteessä A. Lisäksi haastatelluille annettiin lyhyt tietopaketti *Tiedon hyödyntäminen ja avaaminen* -hankkeesta sekä tämän tutkimuksen taustoista.

Puolistrukturoidun haastattelun reliabiliteettia pyritään ylläpitämään ehkäisemällä mahdollisia puolueellisia vastauksia ja vastausten tulkitsemista (Saunders et al. 2019, s. 444). Tässä tutkimuksessa empiirinen laadullinen tieto on kerätty teemahaastattelujen

pohjalta, ja näin ollen kerätty tieto on tapauskohtaista. Haastattelujen vastausten puolueellisuutta on pyritty vähentämään esittämällä haastateltavat tahot anonyymisti, jolloin vastauksien antamisessa voidaan keskittyä suoraan käsiteltävään aihepiiriin ilman ns. mainevaikutuksia.

Laadulliset haastattelutulokset analysoidaan induktiivisesti. Induktiivinen lähestymistapa sopii tutkimuksiin, joissa laajaa mittavaa kerättyä tietoa halutaan tiivistää. Toinen sopiva tilanne on, jos kerätystä tiedosta halutaan luoda yhteyksiä tutkimuksen päämäärään. Tämä on kuitenkin tehtävä ylläpitäen tutkimuksen luotettavuutta. Kolmas tilanne on se, kun kerätystä tiedosta halutaan luoda tieteellinen malli tai teoria (Thomas 2003). Tässä tutkimuksessa induktiivinen lähestymistapa sopii ensimmäisten kahden tilanteen osalta hyvin.

3.1 Tutkimusongelma ja -kysymys

Tiedon hyödyntäminen ja avaaminen –hankkeella on yhtenä tarpeena kehittää kansallista avoimen tiedon hyödyntämistä. Sitä varten on tärkeä löytää ja tunnistaa metatietojen haasteet ja huomioon otettavat asiat avoimen tiedon avaamisessa ja hyödyntämisessä, sekä etsiä parannuskohtia avoimen tiedon metatietojen semanttisessa ja teknisessä yhteentoimivuudessa. Tämä lähtökohta toimii tämän diplomityön tutkimusongelmana.

Diplomityön tutkimuskysymyksenä on:

”Mitä metatietojen haasteita ja huomioon otettavia asioita on olemassa avoimen tiedon avaamisessa ja hyödyntämisessä?”

Tähän tutkimuskysymykseen saadaan vastaus haastattelujen yhteenkokoamisesta, analysoinnista sekä teoriapainotteisesta pohdinnasta. Tavoitteena on luoda uutta tietoa valtiovarainministeriön *Tiedon hyödyntäminen ja avaaminen* -hankkeen yhtä osatavoitetta varten, joka on *”metatietoihin kohdistuvat muutostarpeet on tunnistettu ja dokumentoitu*” (Valtiovarainministeriö 2020).

3.2 Haastateltavat organisaatiot ja henkilöt

Kohdeorganisaatiot on valittu sellaisten julkisten ja julkisomisteisten organisaatioiden joukosta, jotka ovat jo, tai ovat juuri avaamassa tietovarantojaan. Valitut organisaatiot ovat sellaisia, jotka suhtautuvat positiivisesti tietojen avaamiseen ja hyödyntämiseen, jotta löydetyt haasteet soveltuvat mahdollisimman hyvin siihen tilanteeseen, kun tietoja on jo päätetty avata. Täsmennettäköön, että tarkoituksena ei ole siis etsiä ongelmakohtia organisaatioiden halusta jakaa tietoa.

Tutkimuksessa on tarkoitus kerätä tietoa mahdollisimman kattavasti ja yleispätevästi Suomen avoimen tiedon metatietojen haasteista ja huomiosta. Näin ollen haastattelupyynnöt ja haastateltavien organisaatioiden valinnat kohdistuivat Suomen merkittävimpiin tiedon tuottajiin ja tarjoajiin julkisella puolella. Haastateltavat organisaatiot on valittu niin, että kukin organisaatio edustaa eri tietovarantotyyppiä, jotka esiteltiin kappaleessa 2.2. Valintaan vaikutti myös kohdeorganisaatioiden halu osallistua haastatteluun sekä mahdollisten haastateltavien henkilöiden aikataulut.

Haastateltavat henkilöt ovat valituissa organisaatiossa työskenteleviä henkilöitä, joilla on kokemusta tiedon avaamisesta, metatiedoista tai molemmista näistä. Saman organisaation haastateltavista henkilöistä pyritään keräämään tietoa erilaisista näkökulmista ja projektikokemuksista. Haastattelijoita ei ole rajattu roolin tai teknisen tietämyksen perusteella. Haastateltavat organisaatiot ja henkilöt esitetään anonyymina. Tässä tutkimuksessa kerätyt tiedot ovat yleispäteviä, ja kerätty suomalaisilta asiantuntijoilta, joiden henkilöllisyys ja organisaatio ei vaikuta tutkimustuloksiin.

Haastattelujen kohteiksi valikoitiin merkittäviä Suomessa toimivia julkisia toimijoita. Haastatteluja pidettiin kolmen eri organisaation työntekijöille. Organisaatiot esitetään tässä tutkimuksessa anonyymeinä, ja niihin viitataan seuraavilla nimillä: Organisaatio A, Organisaatio B ja Organisaatio C.

Kunkin organisaation päätoiminen avattava tieto voidaan jakaa kappaleessa 2.2 esiteltyihin tietovarantotyyppeihin. Nämä tietovarantotyypit ovat paikkatieto, maan havainnointi ja ympäristötieto, säätieto, tilastotieto, yritys- ja yritysten omistustiedot sekä liikkuvuustiedot. Kaikki kolme organisaatiota suhtautuu hyvin positiivisesti tiedon avaamiseen. Toimijat ovat jo avanneet suurimman osan aineistoistaan ja avaavat aktiivisesti lisää, sekä ylläpitävät ja kehittävät olemassa olevia aineistoja ja niiden rajapintoja.

Organisaatio A on suuri julkinen toimija, joka toimii pääasiassa paikkatiedon parissa. Sillä on valtava määrä erilaisia rekisterietä sekä paikkatieto- ja maan havainnointi- ja ympäristötietoaineistoja, joita se kerää ja ylläpitää aktiivisesti. Lähtökohtaisesti kaikki aineisto tässä organisaatiossa on avointa, mutta tietyin osin avoimuutta rajoittavat GDPR-säädökset. Poikkeuksena ovat myös tietyt hyvin tarkat aineistot, jotka ovat osittain käyttörajoitettuja. Aineistoja on saatavilla suoraan omien rajapintojen kautta, sekä ulkopuolisista dataportaaleista.

Haastatellut henkilöt Organisaatio A:sta ovat olleet jo kauan tekemisissä organisaation tietopalveluiden parissa. He ovat olleet mukana eri tiedon avaamisen projekteissa organisaation sisällä, sekä ovat olleet avaamassa ja tukemassa tiedon avaamista heidän ulkoisille asiakkailleen. Haastateltavat henkilöt ovat toimineet muun muassa organisaation

keskushallinnon ja tietopalveluiden johtotehtävissä sekä tietopalveluiden asiantuntijatehtävissä.

Organisaatio B:n päätietovarantotyyppinä voidaan ajatella olevan tilastollinen tieto, mutta organisaatio toimii myös vahvasti paikkatiedon parissa. Organisaatio B kerää ja ylläpitää tilastotietoa, luokitustietoa, paikkatietoa sekä metatietojen tilastollista tietoa eri aineistoista. Aineistoa kerätään jatkuvasti, ja suurin osa siitä on avoimena saatavilla. Osa aineistoista ei kuitenkaan ole avattuna henkilötietojen vuoksi, ja osa on saatavilla maksua vastaan. Tästä johtuen tiedon myyminen on osa organisaation vuosittaista budjettia. Suurin osa aineistosta on julkisena organisaation omassa dataportaalissa. Joitain aineistoja annetaan suoran rajapinnan kautta, ja pieni osa on jaettuna myös organisaation ulkopuolisiin dataportaaleihin.

Organisaatio B:n haastatellut henkilöt ovat toimineet muun muassa kokonaisarkkitehtuurin, sovelluskehityksen ja tietopalveluarkkitehtuurin parissa. Lisäksi haastateltavista henkilöistä löytyy esimies- ja kehittämispäällikkökokemusta. Henkilöt ovat olleet vahvasti mukana tiedon avaamisessa, metatietojen luokittelussa, avoimen tiedon rajapinnoissa sekä tilastollisessa kehityksessä.

Organisaatio C tekee puolestaan tutkimusta ja luo aineistoja pääasiassa säätietojen muodossa. Aineistoja ylläpidetään ja tietojen keräysprosessissa tehdään myös aktiivista analyysiä, jonka tuloksista viestitään laajasti yhteiskunnalle. Säätietojen lisäksi mukana on myös paikkatietoa, sillä säätieto itsessään sisältää paikkatietoa. Organisaatio C:n aineistojen säätietoihin ja paikkatietoihin on aina sidottuna hyvin vahvasti aikaulottuvuus. Näitä monimutkaisia aineistoja on avattuna pääasiassa organisaation omassa dataportaalissa. Periaatetasolla Organisaatio C avaa ja jakaa kaiken sen keräämän käytettävän tiedon.

Organisaatio C:n haastatellulla henkilöllä on pitkä kokemus asiakastuote- ja sovelluskehityksestä sekä tietoarkkitehtuurista. Tämän lisäksi hän on ollut mukana eri tiedon avaamisen työryhmissä sekä aineistojen avaamisprosessissa. Hän on myös toiminut pääarkkitehtina.

Tämän tutkimuksen haastateltavat organisaatiot ja niiden henkilöt ovat kuvattuna alla olevassa taulukossa. Useimmiten organisaatio toimii useiden eri tietovarantotyyppien kanssa, ja ne ovat kuvattuna taulukossa niin, että ensimmäisenä mainittu tietovarantotyyppi voidaan katsoa olevan organisaation pääasiallinen tietovarantotyyppi. Organisaatioiden haastateltavat henkilöt on nimetty organisaation nimen mukaisesti numerojärjestyksessä niin, että saman organisaation henkilöt omaavat saman alkukirjaimen. Kunkin haastateltavan henkilön kuvaus ja tausta lukevat lyhyesti taulukossa henkilön perässä.

Taulukko 1: Haastattelujen organisaatiojakauma ja henkilöiden kuvaus

Organisaatio	Tietovarantotyyppi	Haastateltava	Kuvaus ja tausta
A	Paikkatieto, maan havainnointi ja ympäristötieto	A1	Johtaja, tietopalvelut. Ollut mukana tiedon avaamisessa pitkään.
		A2	Tietopalvelu asiantuntija. Aineistojen avaaminen ulkoisille asiakkaille.
B	Tilastotieto, paikkatieto	B1	Järjestelmä asiantuntija. Kokonaisarkkitehtuuri ja tietotekniset sovellusprojektit. Avoimen tiedon luokitus rajapinnat.
		B2	Tietopalvelusuunnittelija. Tekniset ja tilastolliset tehtävät. Pitkä kokemus tiedon avaamisessa.
		B3	Kehittämispäällikkö. Internetpalveluiden kehittäminen, verkkopalvelut, julkaisutoiminta ja esimiestehtävät.
C	Säätiedot, paikkatieto	C1	Pääarkkitehti- ja sovelluskehittäjätehtävät. Aktiivinen avoimen tiedon parissa eri ryhmissä.

Taulukosta nähdään, että haastateltavilla henkilöillä on kokemusta ja osaamista avoimesta tiedosta, metatiedoista tai molemmista näistä. Organisaatiot ovat aktiivisesti tekemisissä avoimen tiedon kanssa teknisesti ja hallinnollisesti, joten haastatteluissa ilmi käyviä haasteita sekä huomioon otettavia asioita saatiin kerättyä useista eri näkökulmista.

3.3 Haastattelukysymykset

Haastattelukysymyksillä pyritään keräämään tietoa metatiedon haasteista ja mahdollisista ratkaisuista tiedon tarjoajan näkökulmasta. Lisäksi pyritään keräämään tietoa mahdollisista metatiedon ongelmista avatun tiedon hyödyntämisessä sekä selvitetään mahdollisia eri tietotyyppien ominaispiirteitä tiedon avaamisessa. Haastattelukysymykset ovat nähtävissä liitteessä A.

Haastattelukysymysten peruskysymyksillä saadaan kerättyä haastateltavan organisaation ja henkilön taustatietoa haastattelusta kerättävän tiedon kontekstiksi. Peruskysymysten tiedot toimivat pohjana seuraaviin kysymyksiin, sekä avaavat näkökulmaa, josta avointa tietoa ja metatietoja käsitellään.

Kysymyksellä *”Mitä metatietojen haasteita organisaatiolla on tullut vastaan tiedon avaamisessa?”* ja sen alakysymyksillä kerätään tietoa mahdollisesti vastaan tulleista haasteista, joita tiedon tarjoaja on kohdannut tietoa avattaessa. Kysymykset toimivat haasteen kartoittamisen alustana, jonka jälkeen esiin nostettuja haasteita ja huomioita tarkennetaan jatkokysymyksillä.

Viimeisellä kysymyksellä *”Mitä metatietojen haasteita tiedon hyödyntämisessä on tullut vastaan?”* kerätään tietoa mahdollisista haasteista, joita organisaatiolle tai muille tiedon hyödyntäjille on tullut vastaan avattua tietoa hyödynnettäessä. Näistä haasteista on voinut tulla tietoa organisaatiolta itseltään, tai esimerkiksi muiden organisaatioiden palvelupyyntöjen tai kysymysten kautta.

Haasteista ja huomiosta kerätään tietoja sen mukaisesti, että niiden syyt ja seuraukset tulevat hyvin selville. Kerätyt haasteet ja huomiot tuodaan yhteen pohdintaa varten, jonka jälkeen luodaan kattava kuva avoimen tiedon metatietojen haasteista ja huomioista.

3.4 Haastattelujen kulku

Haastattelut pidetään noin tunnin mittaisina puolistrukturoituina teemahaastatteluina. Niiden sisältämät kysymykset ovat rakenteeltaan avoimia. Haastattelujen aikana on tarkoitus selvittää mitä haasteita ja ongelmia organisaatiolla on tullut vastaan tietoja avattaessa, miten ne ratkaistiin, sekä mitä on otettava huomioon eri ongelmatilanteissa. Lisäksi selvitetään, jos organisaatioille on tullut tietoa avatun tiedon hyödyntämisen haasteista. Haastatteluissa pyritään myös huomioimaan mitä erityispiirteitä kullakin kohdeorganisaatiolla on tiedon avaamisen suhteen, ja mitä erityispiirteitä kyseinen avattu tietovarantotyyppi on tuonut eri tilainteisiin.

Tutkimuksen haastattelut sovittiin erikseen osapuolille sopiviin ajankohtiin, ja ne pidettiin etäyhteyksillä. Haastateltaville lähetettiin etukäteen liitteen A kysymyspatteristo ja lyhyt kuvaus tutkimuksen aihepiiristä luettavaksi. Haastateltavien ei tarvinnut haalia mitään etukäteistietoja haastatteluja varten. Kukin haastattelu oli noin tunnin mittainen puolistrukturoitu haastattelu. Haastattelun jälkeen mahdollisia vastausten tarkennuksia tehtiin sähköpostitse ja jatkohaastatteluilla. Pidetyissä haastatteluissa syntyi paljon uutta tietoa avoimesta tiedosta, metatiedoista ja niiden haasteista. Kerätyistä tiedoista on jä-

sennelty ja koottu tähän tutkimukseen merkittävimmät havainnot. Tutkimuksessa suoritetut haastattelut, haasteet ja huomioonotettavat asiat, sekä niiden kytkeytyminen toisiinsa havainnollistetaan taulukoilla ja kuvilla.

Haastateltavat organisaatiot ja henkilöt esitetään tässä tutkimuksessa anonyymeinä, sillä tutkimuksen näkökulmasta haastateltavien tahojen identiteetti ei vaikuta saatuihin tuloksiin. Lisäksi joissain haastatteluissa kävi ilmi, että avoimen tiedon metatietojen haasteita oli helpompi käsitellä kriittisestä näkökulmasta anonyymisti. Samalla haastatteluissa kyettiin keskittymään yleisesti tutkimuksen aihepiiriin ilman organisaatioiden markkinoimista.

Puolistrukturoidut teemahaastattelut suoritettiin seuraamalla liitteen A haastattelukysymyksiä. Tarkentavia jatkokysymyksiä käsiteltiin myös aktiivisesti. Haastatteluissa esiin nousseita avoimen tiedon metatietoihin liittyviä haasteita ja huomioon otettavia asioita käsiteltiin niin kauan, että niiden perimmäiset syy-seuraussuhteet tulivat selviksi. Haasteita ja huomioon otettavia asioita pyrittiin löytämään niin tiedon jakajan, kuin aineiston käyttäjän näkökulmasta. Osa haasteista linkittyi suoraan avoimen tiedon metatietoihin, kun taas osa koski välillisesti metatietoihin avoimen tiedon käytettävyyden, löydettävyyden ja luotettavuuden kautta.

Haastatteluja suoritettiin yhteensä kuusi kappaletta, ja ne kaikki suoritettiin kevään 2021 aikana. Haastatteluja saatiin kolmesta eri organisaatiosta. Suoritettujen haastattelujen lisäksi haastattelupyyntöjä lähetettiin useisiin eri organisaatioihin, joista osa kieltäytyi ja osa jätti vastaamatta. Kuusi suoritettua haastattelua onnistui hyvin. Näistä saatiin paljon uutta tietoa avoimen tiedon metatiedoista. Varsinaisia tutkimuskysymyksen aihepiirin metatietojen haasteita ja huomioon otettavia asioita saatiin kerättyä 17 kappaletta, jotka esitetään tarkemmin tässä tutkimuksessa.

3.5 Haastattelujen analysointi

Suoritetut haastattelut nauhoitettiin analysointia varten. Haastattelujen vastaukset analysoidaan induktiivisella sisällön analyysilla, jossa pyritään luomaan kuva tutkittavasta ilmiöstä tiivistetyssä muodossa (Kyngäs & Vanhanen 1999). Sisällön analyysi toimii hyvin strukturoimattoman aineiston, esimerkiksi dialogien tai haastattelujen analysoimiseen (Weber 1985).

Induktiiviselle analysointiprosessille ei ole olemassa yhtä yleispätevää ohjeistusta. Induktiivinen analysointiprosessi tehdään aineisto lähtöisesti (Sandelowski 1995), ja tässä tutkimuksessa analysoitavana aineistona toimii haastattelujen taltioinnit. Induktiivisessa

analysoinnissa on kolme vaihetta, jotka ovat pelkistäminen, ryhmittely sekä käsitteellistäminen (Kyngäs & Vanhanen 1999). Pelkistämisessä aineistolta kysytään tutkimuksen mukaisia kysymyksiä ja kirjataan aineistosta relevantit vastaukset mahdollisimman tarkasti ylös. Seuraavaksi pelkistetyt vastaukset ryhmitellään sopivien tulkinnan mukaisten kategorioiden alle. Lopuksi luotuja kategorioita käsitteellistetään, purkamalla niitä pienempiin osiin tai yhdistämällä toisiinsa, kunnes aineiston vastaukset on saatu jaettua sopiviin ryhmiinsä (Dey 1993).

Tässä tutkimuksessa induktiivinen sisällön analyysi suoritettiin tunnistamalla suoritettujen haastattelujen vastauksista esiin tulleita avoimen tiedon metatietojen haasteita pelkistämällä vastauksista haasteiden kuvaukset sekä niiden merkitykset ja seuraukset. Pelkistetyt vastaukset kategorisoitiin sopivien ryhmien alle niin, että samaan haasteeseen koskevat vastaukset olivat sopivissa ryhmissä. Näitä tunnistettujen haasteiden kuvauksia ja haastattelujen vastauksia esitetään seuraavassa luvussa.

4. TULOKSET JA NIIDEN TARKASTELU

Tässä luvussa käsitellään suoritettujen teemahaastattelujen relevantit tulokset. Tapauskohtaiset haastatteluissa esiin tulleet haasteet ja huomiot, jotka koskettavat tämän tutkimuksen aihepiiriä, on kerätty haastattelujen vastauksista. jotka ovat käsitelty induktiivisella sisällön analyysilla, joka käsiteltiin kappaleessa 3.5. Haastattelujen perusteella on tunnistettu 17 erillistä avoimen tiedon metatietojen haastetta ja huomioon otettavaa asiaa, jotka ovat linkittyneet toisiinsa. Nämä haasteet ja huomioon otettavat asiat esitetään havainnollisessa syy-seuraussuhteita kuvaavassa verkossa.

4.1 Haastattelujen tulokset

Haastatteluista saatiin kerättyä paljon uutta tietoa muun muassa organisaatioiden yleisestä toiminnasta, tiedon avaamisen käytännöistä, henkilöiden taustoista sekä erilaisten projektien kuluista teknisestä ja hallinnollisesta näkökulmasta. Tähän tutkimukseen on tuote esille vain ne haastattelujen löydökset, jotka kuuluvat suoraan tutkimuksen aihepiiriin tiedon avaamisen metatietojen haasteista ja huomioista.

Tutkimuksessa tunnistetut avoimen tiedon metatietojen haasteet ja huomioon otettavat asiat luotiin yhdistämällä suoritetuissa haastatteluissa esiin tulleita huomioita. Haastattelujen vastauksista muodostettiin 17 haastetta induktiivisella lähestymistavalla. Haastattelujen vastauksia ensin tiivistettiin ja pyrittiin erottelemaan eri haasteita ja huomioita. Nämä tulkittiin tapauskohtaisesti ja pyrittiin luomaan yhteyksiä tämän tutkimuksen päämäärään induktiivisesti yhdistelemällä kaikkien haastattelujen vastauksia.

Tunnistetut avoimen tiedon metatietojen haasteet ja huomioon otettavat asiat sekä niiden yhteydet esitetään erikseen. Näitä perustellaan haastattelujen vastauksien suorilla lainauksilla, jotka vahvistavat tunnistettuja haasteita ja lisäävät tutkimuksen luotettavuutta. Vastauksien suorat lainaukset esitetään edellisen luvun nimityksillä, jotka esitettiin taulukossa 1.

Kaikki haastatteluissa esiin nousseet huomiot ovat yhtä arvokkaita, vaikka osa huomioista nousi esiin enemmän kuin kerran. Avoimen tiedon metatietojen haasteita koskevissa huomioissa oli nähtävillä selvät syyt ja niiden vaikutukset, jotka ovat kuvattuna seuraavassa taulukossa. Löydetyt 17 haastetta ja huomiota on nimetty kirjaimilla A:sta Q:hun.

Taulukko 2: Haastatteluissa ilmenneet haasteet ja huomiot (jatkuu seuraavalle sivulle)

Huomio / haaste	Syy	Vaikutus
A. Aineiston jalostusasteen epäselvyys	Metatiedoissa ei ilmene mitä mahdollisia muokkauksia aineiston datalle on jo tehty.	Aineiston ymmärtäminen ja käytettävyys kärsii.
B. Metatietojen semanttisuuden puute	Aineisto julkaistaan ilman riittävän tasoista kuvausta ja metatiedot eivät itsestään selitä tarpeeksi aineistosta.	Semanttisessa mielessä metatiedot eivät kerro mistä, miten ja milloin data on kerätty.
C. Aineiston kyseenalaistamisen mahdollisuuden puute	Aineisto kuvaillaan liian suppeasti yhdestä näkökulmasta.	Käyttäjillä voi olla tietämättään omaan käyttötarkoitukseensa vääränlaista tietoa ja aineiston luotettavuus ei ole taattu.
D. Aineiston valtava koko	Valtavia aineistoja otetaan käyttöön osina, joissa metatiedot eivät pysy mukana.	Metatietojen ja aineiston luotettavuus kärsii.
E. Aineiston heikko luotettavuus	Aineiston metatiedot eivät kuvaa tarpeeksi kattavasti ja tarkasti, miten, mistä ja milloin data on luotu.	Loppukäyttäjät ja kehittäjät eivät tiedä aineiston luomisesta ja käyttämisestä.
F. Aineiston heikko hyödynnettävyys	Aineisto on vaikeasti löydettävä ja sen metatiedot eivät kuvaa tarpeeksi hyvin mitä aineisto sisältää.	Aineisto ei tule käyttöön toivotulla laajuudella.
G. Metatiedon heikko laatu	Metatieto ei vastaa laatukriteerejä.	Epälaadukas metatieto heikentää aineiston käytettävyyttä.
H. Aineiston heikko hakukonelöydettävyys	Huono semanttinen metatieto heikentää hakukoneiden mahdollisuutta löytää aineisto hakusanoilla.	Aineiston potentiaali ei tule laajasti hyödynnettäväksi.
I. Eri kielien käyttäminen metatiedoissa	Metatiedoissa on epäjohdonmukaisesti eri kieliä. Data ja kuvaukset sisältävät eri kielistä tekstiä.	Metatietojen johdonmukaisuus ja semanttisuus kärsivät.

J. Tietotyyppien spesifiset metatiedot	Tietyt tietotyypit sisältävät hyvin spesifistä metatietoa, jotka ovat välttämättömiä vain tietyille aineistoille.	Yleiset dataportaalit eivät taivu tiedon avaamiseen. On luotava oma portaali.
K. Metatietojen epäsopiva määrä	Aineistojen metatiedon määrän vaatimukset ovat liiallisia tai liian vähäisiä.	Metatieto jää vähäiseksi tai sitä on epätarkkana liian paljon.
L. Metatietojen pysyvyyden puute	Aineistojen siirtyminen rajapinnan läpi. Aineiston ja metatiedon synkronoinnin puute.	Metatietojen ajantasaisuus ja aineiston käytettävyys kärsivät.
M. Metatietojen ajantasaisuuden puute	Aineistojen valtava koko ja jatkuvuuden ja valvonnan puute tiedon avaajan näkökulmasta.	Epäajantasainen metatieto ei enää kuvaa aineistoa riittävällä tasolla ja aineiston luotettavuus kärsii.
N. Metatietojen semanttiset ongelmat ohjelmissa	Ohjelmat eivät osaa tulkita semanttista tietoa ja aineistojen semanttisuus ei ole universaalia.	Semanttisten tietojen sisältämät asiat jäävät huomiotta. Tiedon merkitys kärsii.
O. Yhteisten käytäntöjen puuttuminen metatiedoista	Yleisten metatietojen määrittelevien käytäntöjen ja standardien puuttuminen tai ne ovat epätarkkoja.	Aineistojen ymmärtäminen ja hyödyntäminen on haasteellista.
P. Liian korkeatasoinen metatieto	Avattavan tiedon selkeän universaalin infrastruktuurin puuttuminen.	Metatieto ei ole eksaktia ja tiedon hyödyntäjillä ei ole tarkkaa tietoa aineistojen sisällöstä.
Q. Liian tekniset metatiedot	Aineistot ja metatiedot luodaan ammattilaisten käyttöön ja vaativat teknillistä ymmärrystä.	Aineiston loppukäyttäjä ei tiedä mitä tietoja hänellä on käytettävissä.

Todettakoon, että haastateltavien organisaatioiden aineistot ovat heille elintärkeitä, mistä johtuen niihin panostetaan valtavasti. Kaikilla valituilla organisaatioilla on paljon kokemusta aineistojen luomisesta, ylläpitämisestä ja kehityksestä. Aineistojen hallinta ja avoin tieto on organisaatioiden toiminnan ytimessä.

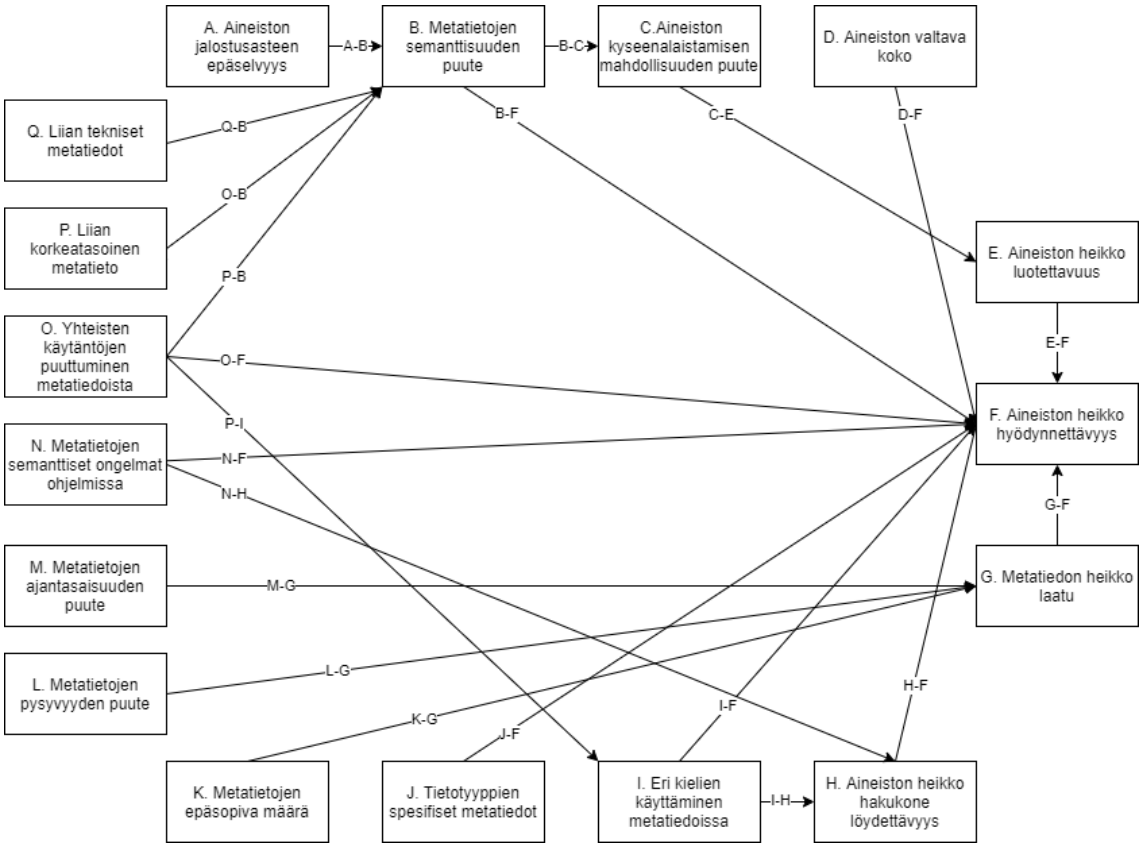
”Aineistot ovat meille elinehto, olemassaolon tarkoitus ” (C1)

Haastatellut organisaatiot jakavat tietojaan yhteiskunnallisista ja laillisista syistä. Tiedon avaamisen voidaan nähdä olevan julkisten organisaatioiden velvollisuus, joka voi tuottaa paljon arvoa eri valtiovallantoimijoille, yksityisille organisaatioille ja kolmannelle sektorille. Tiedon avaamisesta on syntynyt jo nyt uutta liiketoimintaa ja yhteiskunnallista arvoa monenlaisista eri aineistoista.

”Lähtökohtaisesti mitään tietoja ei kerätä pelkästään omaan käyttöön. Tarkoituksena on kerätä ja ylläpitää aineistoja yhteiskunnallista tarvetta varten.” (A1)

Organisaatioiden avaamattomia aineistoja ei ole avattu muutamista syistä. Usein ne sisältävät sellaista henkilötietoa, joka estää tietojen avaamisen. Tietyt aineistot voivat sisältää sellaista tietoa, joka voi pahimmillaan aiheuttaa yhteiskunnallista harmia tiedon avaamisen seurauksena. Joskus taas tietoa ei ole avattu siitä syystä, että sillä ei nähdä olevan käyttötarkoitusta, jolloin tiedon avaamisen vaivaa ei nähdä tai koeta tarpeelliseksi.

Löydettyjen haasteiden ja huomioiden syiden vaikutukset voidaan nähdä liittyvän toisiinsa. Osalla löydetyistä huomioista on samanlaisia vaikutuksia, ja joidenkin haasteiden aiheutuksen syinä ovat samat asiat. Esimerkiksi haasteiden Q ja P voidaan nähdä vaikuttavan haasteeseen B. Näin yhdistämällä löydettyjä haasteita ja huomioon otettavia asioita toisiinsa, voidaan luoda yhteyksille verkko, joka kuvaa verkossa olevien huomioiden loogisia syy-seuraussuhteita.



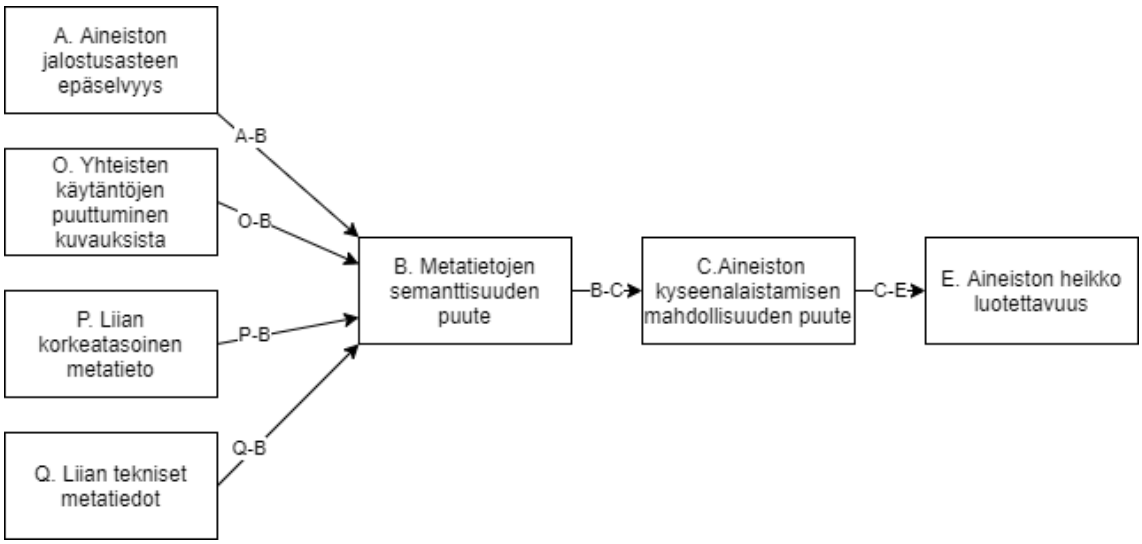
Kuva 1: Haastatteluissa esiin tulleiden haasteiden ja huomioiden verkko

Yllä kuvattuun verkkoon on koottu kaikki haastatteluissa esiin tulleet relevantit haasteet avoimen tiedon metatietojen näkökulmasta. Verkossa käytetään samaa haasteiden kirjainnimitystä kuin taulukossa 2. Haasteet on laitettu kehälle aakkosjärjestyksessä kiertämään kehää myötäpäivään alkaen vasemmasta yläkulmasta. Haasteiden syy-seuraussuhteita on kuvattu nuoliviivoilla alkaen siitä haasteesta, jonka vaikutuksena nuolen loppupään haaste tapahtuu. Nämä loogiset suhteet on nimetty haasteiden kirjainnimitysten mukaisesti, esimerkiksi haasteen N kahta vaikutusta haasteisiin F ja H kuvataan suhteilla N-F ja N-H.

Verkon analysoimista helpottamiseksi verkko on jaettu loogisten suhteiden mukaisesti neljään osaan, jotka on nimetty niiden kokoavan haasteen mukaisesti. Nämä neljä verkon osaa ovat aineiston heikko luotettavuus, metatiedon heikko laatu, aineiston heikko hakukonelöydettävyys sekä aineiston heikko käytettävyys. Verkon neljän osa-alueen sisältämät haasteet ja huomioon otettavat asiat esitetään yksitellen. Lisäksi kuvataan niiden loogisia syy-seuraussuhteita toisiinsa.

4.1.1 Aineiston luotettavuus käyttäjien näkökulmasta

Ensimmäinen haastattelujen tuloksista luotu verkon käsiteltävä osa on aineiston heikko luotettavuus, joka nousi esiin useassa haastattelussa useiden eri syistä. Tässä kokonaisuudessa on mukana seitsemän eri avoimen tiedon metatietojen haastetta ja huomioon otettava asiaa, jotka esitetään johdonmukaisessa järjestyksestä syy-seuraussuhteiden alkupäästä niiden loppu päähän.



Kuva 2: Syy-seuraussuhteet aineiston heikkoon luotettavuuteen

Ensimmäinen käsiteltävä avoimen tiedon metatiedon haasteista ja huomioon otettavista asioista on *A. Aineiston jalostusasteen epäselvyys*. Tällä tarkoitetaan sitä tilannetta, kun aineiston tekniset ja semanttiset metatiedot eivät kuvaa tarpeeksi hyvin mitä aineistolle on tehty ennen sen avaamista. Tämä on hyvin merkittävä haaste etenkin tilastollisen tiedon hyödyntämisessä, jossa datan aggregointia, eli tilastollista muokkaamista sopivampaan muotoon, tehdään paljon.

”Raakadataa voi käsitellä hyvin monin eri tavoin, mutta aggregoidussa datassa, jossa otetaan mukaan yleistyksiä ja yksilöryhmiä, on otettava mukaan vertailukelpoisuus. Esimerkiksi vertailukelpoisuus ajassa ja erinäisissä luokissa, joissa eri joukkoja jaetaan osiin, on tärkeä huomioida. Näissä on otettava mukaan kuvailut aggregointimenetelmistä ja datan muodostuksesta.” (B2)

Haaste A aiheuttaa aineiston kuvailussa semanttisia ongelmia, joka johtaa haasteeseen *B. Metatietojen semanttisuuden puute*. Tätä yhteyttä kuvataan edellisessä kuvassa nuoliviivalla A-B. Haaste B koskee avoimen tiedon ongelmia, jotka johtuvat metatietojen semanttisuuden puutteesta. Metatietojen semanttisuudella tarkoitetaan tässä yhteydessä metatietojen kykyä selittää ja kuvata aineiston tietosisältöä ja käyttötarkoitusta. Näiden lisäksi se sisältää mahdollisia aineiston ohjeistuksia ja alkuperän kuvauksia sekä tietoa aineiston muokkauksista. Haaste B aiheuttaa moninaisia ongelmia, etenkin aineiston hyödyntäjien näkökulmasta. Se vaikeuttaa selvittämään aineistojen käyttömahdollisuuksia ja löydettävyyttä. Lisäksi se hankaloittaa aineistojen hyödynnettävyyttä, jota käsitellään myöhemmin suhteen B-F määrittelyssä.

”On maita, joissa tekstimuotoisten tilastojulkistusten tekeminen on vähäisempää kuin Suomessa. Jos me ei tehtäisi tekstimuotoista tiivistelmää vaan pelkkä tietokantajulkistus, eli julkaistaisiin pelkät taulukot, mutta ei mitään tekstimuotoista selostusta, niin ammattikäyttäjät kyllä tietävät mitä tietoa on saatavilla, mutta muiden käyttäjien tiedon hahmottaminen ja löytäminen on hankalampaa.” (B3)

Tekstimuotoisen metatiedon julkaiseminen on epäteknisille käyttäjille hyvin arvokasta, erityisesti sen takia, että ammattimaisten datan tulkitsijoiden määrä on vähentynyt. Tämän takia tekstimuotoinen tieto on yhä arvokkaampaa tänä päivänä. Esimerkiksi media hyötyy suuresti siitä, että toimittajat voivat tehdä tehokkaasti analyysejä erinäisten julkaisujen semanttisista metatiedoista.

”Metatiedoissa olisi hyvä olla monta eri tasoa, jotta sen löytää ja sitä kuvaillaan tarpeeksi. Siitä on löydyttävä tarvittavat tiedot sen käyttöä varten.” (A1)

Haaste *O. Yhteisen käytäntöjen puuttuminen metatiedoist*a, on ilmeinen haastatteluissa toistuvasti esiintynyt haaste. Tällä tarkoitetaan selkeiden universaalien pelisääntöjen puuttumista, kun käsitellään avattavien aineistojen metatietoja. Tietyn tyyppisten aineistojen metatiedoissa on olemassa yleisesti käytettyjä standardeja ja viitekehyksiä. Esimerkiksi paikkatietoihin spesifioitunut Inspire -viitekehys on kehitetty satojen eurooppalaisten asiantuntijoiden toimesta luomaan yhteiset käytännöt paikkatietojen avaamiselle ja jakamiselle Euroopan julkisille toimijoille. Se pitää sisällään toimintaperiaatteiden lisäksi standardit avattavan tiedon metatiedoille. (European Comission 2021).

”Datan avaamisessa Inspire on ollut tosi hyvä, ja moni on kiittänyt standardipohjaisen viitekehyksen käyttöä, mutta sen monimutkaisuutta monet hyvällä syyllä kritisoivat. Se on iso haaste, mutta aikanaan on valittu se linja, että noudatetaan standardeja ja viitekehyksiä, ja siitä pidetään edelleenkin kiinni, mutta tehdään töitä sen eteen, että viitekehyksen kokonaisuuden käyttöä saataisiin järkevämmäksi, ja on tällä hetkellä hyvää vauhtia menemässä siihen suuntaan.” (C1)

Vielä 2010-luvun alussa standardien käyttäminen ei ollut kaikkien aineistotyyppien keskuudessa välttämätöntä, mutta nykymaailmassa standardien noudattaminen on ehdoton vaatimus tiedon avaamisessa. Standardeja ja yleisiä viitekehyksiä ei aina noudateta orjallisesti absoluuttisella tarkkuudella, vaan ne on mahdollista muokata omaan käyttöön sopivaksi. Tämä tapahtuu kuitenkin niin, että määritetyn standardin perusteet ovat käytössä. Esimerkiksi GSIM (Generic Statistical Information Model) on kansainvälinen aineistojen määritelmiä, ominaisuuksia ja yhteyksiä kuvaava viitekehys, joka voidaan implementoida eri organisaatioiden ja aineistojen käyttöön sopivalla tavalla (UNECE 2021).

Avoimet tiedon standardit ja viitekehykset palvelevat tiedon avaajien lisäksi tiedon hyödyntäjiä. Selkeät yleiset käytännöt helpottavat kehittäjiä ja tiedon hyödyntäjiä saamaan aineiston hyötypotentiaalista arvoa nopeammin ja luotettavammin. Standardit ehkäisevät tiedon hyödyntämisen virheitä, kun esimerkiksi aineistossa käytetyt yksiköt ja määreet ovat tietyn standardin mukaisesti määritelty. Lisäksi standardien ja viitekehysten avulla tiedon avaajilla ja hyödyntäjillä on mahdollisuus tarkistaa ja lukea standardin tarkoista säädöksistä, jolloin mahdollisten epäselvyyksien ratkaiseminen tehostuu.

Yhteisten käytäntöjen puuttuminen voi aiheuttaa itsessään haasteita. On kuitenkin mahdollista, että standardeja ja viitekehyksiä käyttäessäkin voi aiheutua metatietojen semanttisia ongelmia. Näitä syntyy esimerkiksi, jos standardit ja viitekehykset eivät ota semantiikkaa tarpeeksi huomioon. Tätä yhteyttä kuvataan kuvassa 2 nuoliviivalla O-B. Standardeja ja viitekehyksiä noudattaessa metatiedon laatu ja aineiston tekninen kuvaaminen on todennäköisesti hyvällä tasolla, mutta usein ne eivät ota kantaa aineistojen tekstimuotiseen kuvaamiseen. Näin ollen semanttinen metatieto voi jäädä vajavaiseksi, ellei sitä käytetyn tiedon avaamisen viitekehyksessä määritellä.

Yleisesti käytetyt avoimen tiedon standardit ja viitekehykset voivat aiheuttaa suoraan tai välillisesti ongelmia metatietojen tekniseen tai kuvailevaan tarkkuustasoon. Tätä tilan-

netta kuvataan haasteella *P. Liian korkeatasoinen metatieto.* Standardit toimivat teknisessä mielessä hyvin ja aineistolle tuttujen ammattilaisten käytössä niiden käyttäminen ei aiheuta isompia haasteita. Kuitenkin tiedon avaamisessa tiedon hyödyntäjiä on monenlaisia, jolloin liian korkeatasoiset ja epätarkat metatiedot heikentävät tiedon ymmärrettävyyttä, esimerkiksi sellaisten kehittäjien keskuudessa, joille kyseinen aineisto ei ole ennalta tuttu.

”Ongelmana on se, että metatieto ei ole riittävän eksaktia, että sen perusteella kehittäjät voisivat tehdä mitään älykkäitä hakuja tai selvittää mitä pitää hakea.” (C1)

Haaste P aiheuttaa sellaisia tilanteita aineiston käyttäjille, että he saavat metatiedoista liian yleistasoisen kuvan aineistosta. Metatieto voi esimerkiksi säätiedoissa osoittaa ainoastaan sen tiedon, että Suomessa mitataan lämpötilaa, vaikka todellisuudessa pitäisi saada tietoon, että miltä asemalta pitää hakea tietoa, jos haluaan säähavaintotietoa tietystä alueesta tiettyyn aikaan. Näin haasteen P osaseurauksena tapahtuu haaste B, jota kuvataan nuoliviivalla O-B kuvassa 2.

Viimeinen haastatteluissa esiin tullut haasteen B osasyynä oleva haaste on *Q. Liian tekniset metatiedot.* Tämä haaste on määritelmältään lähellä edellistä haastetta P, mutta se ottaa kantaa teknisen ja semanttisen metatiedon teknisyyden tasoon.

”Jos tiedot halutaan kunnolla hyötykäyttöön, niin metatiedot pitäisi näkyä loppukäyttäjälle asti. Näin ollessa käyttäjän olisi hyvä tietää mitä mikäkin tieto on.” (A2)

Metatiedon semanttisuuden puute aiheuttaa suoraan sen, että avoimen tiedon käyttäjä ei tiedä mitä tietoa hänellä on käytössään. Tämä vaikeuttaa aineiston kyseenalaistamisen mahdollisuutta ja arviota siitä, ratkaiseeko kyseisen aineiston käyttäminen sen ongelman, johon käyttäjä pyrkii vastaamaan. Tätä tilannetta kuvataan haasteella *C. Aineiston kyseenalaistamisen mahdollisuuden puute* ja nuoliviivalla B-C.

”Tietoa ei saa käyttää suoraan, vaan sitä on kyseenalaistettava. Aika usein jää maalaisjärjen varaan kysyä miten mittarit ja luvut on oikeasti laskettu. Jos sitä ei ymmärrä, niin ei voida ymmärtää sitä mihin dataa voidaan käyttää. Tämä on haastavaa, mutta

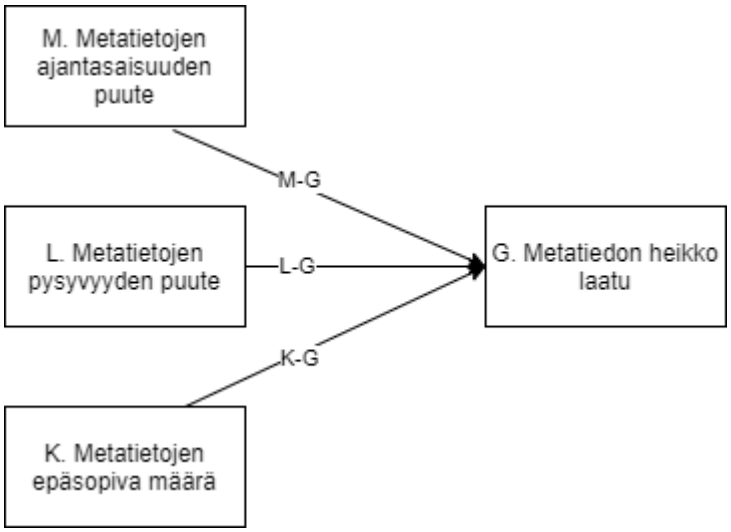
tärkeää. Luvut eivät ole vain lukuja, vaan niiden konteksti ja hyödyntämisen tiedot ovat ytimessä." (B2)

Usein aineistolla on paljon rajoitteita, jotka eivät käy ilmi ilman suoria semanttisia selostuksia. Etenkin avoimessa tilastollisessa tiedossa on omattava sopiva tilastolukutaito tai kyky kyseenalaistaa aineiston käyttömahdollisuuksia semanttisten kuvausten kautta. Aineiston sisältö on oltava selvillä sen käyttäjille, jotta sen hyötypotentiaali tulee käyttöön oikealla tavalla. Tämä vaikuttaa suoraan tämän verkon osa-alueen viimeiseen haasteeseen *E. Aineiston heikko luotettavuus*, joka kuvattiin edellisessä kuvassa nuoliviivalla C-E.

Haaste E kuvaa yleisesti kaikkia haastatteluissa esiin tulleita haasteita ja huomioon otettavia asioita aineiston luotettavuudesta. Se kattaa edellä mainittujen haasteiden välilliset seuraukset liittyen aineiston käyttämiseen, sen alkuperään, jalostukseen, kuvauksiin ja käytettyihin viitekehyksiin sekä standardeihin. Näistä aiheutuvien haasteiden lopputuloksena aiheutuu tilanne, jossa aineiston käyttäjä ei tiedä mitä tietoa hänellä on käytössään, ja mihin tarkoitukseen sitä voi hyödyntää. Nämä epäselvyydet tekevät aineiston käyttämisestä epäluotettavaa.

4.1.2 Metatiedon heikko laatu

Seuraava käsiteltävä verkon osa on metatiedon heikko laatu, joka pitää sisällään neljä metatietojen haastetta ja huomioon otettavaa asiaa. Verkon osan kokoava haaste *G. Metatiedon heikko laatu* tarkoittaa tilannetta, jossa metatieto ei kuvaa aineistoa riittävällä tavalla. Metatieto on tietoa tiedosta. Sen laadun heikentyessä metatieto ei enää anna aineistosta totuudenmukaista tietoa. Metatieto voi olla myös tämän tutkimuksen teoriaosuudessa mainittujen metatiedon laatukriteerien vastaista.



Kuva 3: Syy-seuraussuhteet metatiedon heikkoon laatuun

Ensimmäinen käsiteltävä haaste, joka johtaa heikkoon metatiedon laatuun, on haaste *M. Metatietojen ajantasaisuuden puute*. Avoin tieto ei aina pidä sisällään vain staattisia aineistoja, vaan ne voivat muuttua ja kasvaa. Näissä tapauksissa aineiston metatiedon pitäisi pysyä ajantasaisena ja kuvata aineistoa jatkuvasti oikein.

”Metatietoihin liittyen iso haaste on saada metatiedot pysymään ajantasaisena. Tietomassojen on käytännössä pakko olla puoliautomaattisia tai osin täysin automaattisia, jotta olisi mahdollista pitää metatiedot ajan tasalla. Tämä on tuottanut paljon työtä ja siinä on silti vielä ongelmia.” (C1)

Metatiedon ajantasaisuuden puute aiheuttaa sen, että aineistoa ei kuvata enää todenmukaisella tavalla, kun aineisto muuttuu esimerkiksi sen kasvaessa. Metatiedon epäajantasaisuus aiheuttaa sen, että metatiedon oikeellisuus kärsii. Tätä metatiedon laatuun vaikuttavaa syy-seuraussuhdetta on kuvattu nuoliviivalla M-G.

Seuraava avoimen tiedon käsiteltävä haaste on *L. Metatietojen pysyvyyden puute*, jolla kuvataan metatietojen muuttumista virheelliseksi siirtyessä eri rajapintojen läpi tai metatietojen sisältämien tietojen poistumista. Metatietojen ylläpidossa on iso työ, ja se aiheuttaa usein paljon käsin tehtävää työtä. Sitä on pyritty kehittämään esimerkiksi niin, että aineistojen kuvaukset päivittyisivät automaattisesti kaikille käytetyille jakelukanaville eri rajapintojen läpi.

”Aineistot voivat olla tarjolla monta eri kanavaa pitkin eri rajapintapalveluissa, jossa voidaan kohdekohtaisesti pyytää tiettyä dataa. Kuvausten on tultava näissä kanavissa mukana. Matkan varrella ollut useita sellaisia tapauksia, että metatietoja piti uudistaa sellaiseksi, ettei niitä tarvinnut kuvata uudelleen useassa eri paikassa moneen kertaan.” (A1)

Metatietojen pysyvyys voi kärsiä myös silloin, kun metatietojen sisältämät linkit lopettavat toimimasta. Näin voi tapahtua esimerkiksi silloin, jos jonkin aineiston kuvauksia tai muita metatietoja on liitetty aineistoon URL-linkkien avulla, joiden toimiminen loppuu epätarkoituksenmukaisesti.

Haaste *K. Metatietojen epäsopiva määrä* kuvaa tilanteita, kun aineisto sisältää liikaa tai liian vähän tarvittavia metatietoja. Metatietojen liian vähäinen määrä heikentää suoraan metatietojen laatua, jos tietoa tiedosta tuotetaan liian vähän. Näin metatietojen täydellisyys on heikkoa. Tätä yhteyttä on kuvattu nuoliviivalla K-G.

”Osa käyttäjistä haluaa kaiken mahdollisen tiedon, ja osa ei halua sitä liikaa. Käyttäjästä riippuu mitä halutaan ja kaikkien käyttäjien tarpeisiin on vaikea vastata.” (B1)

Avoimen tiedon metatietojen liian suuri määrä ei itsessään ole huono asia, mutta se voi aiheuttaa sen, että käyttäjä ei käytä tarvittavaa aikaa niiden tutkimiseen. Näin metatietojen epäsopiva määrä voi aiheuttaa sen, että avoimen tiedon hyötypotentiaali jää joissakin tapauksessa käyttämättä, tai aineistoa käytetään virheellisesti, kun metatieto tulkitaan väärin.

4.1.3 Aineiston heikko hakukonelöydettävyys

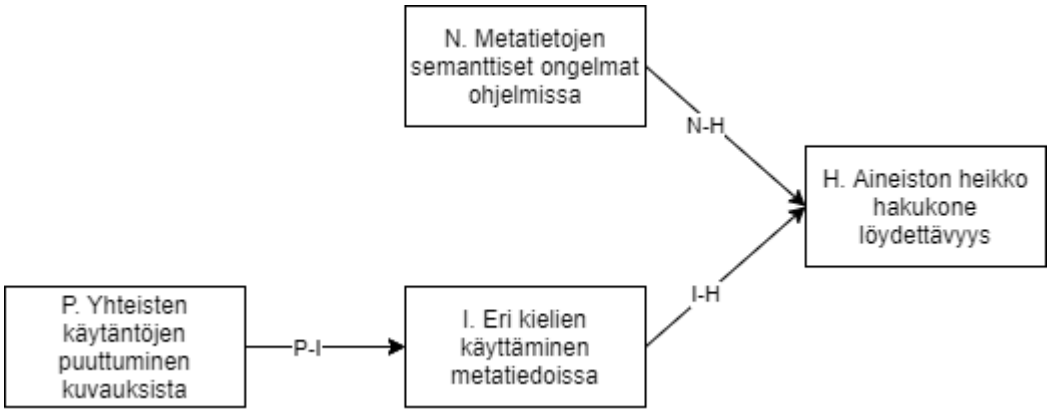
Aineiston hyödyntämiseen vaaditaan jonkinasteista teknistä osaamista ja ymmärrystä aineiston käyttämisestä, jota voi saada kattavista metatiedoista. Kuitenkaan se ei aina riitä, vaan ensimmäinen askel avoimien aineistojen arvonluonnista syntyy niiden löytämisestä. Aineistojen hyötypotentiaali ei tule käyttöön, jos mahdolliset hyödyntäjät eivät löydä niitä. Tästä syystä aineiston löydettävyys voidaan nähdä olevan hyvin merkittävä tekijä, ja metatiedoilla on oma osuutensa sen parantamiseen.

”Huono tilanne on se, että käyttäjät eivät löydä aineistoa huonon metatiedon takia. Silloin koko hyöty on menetetty ja arvoa ei synny.” (A1)

Tätä tapahtumaa kuvataan seuraavassa kuvassa haasteella *H. Aineiston heikko hakukonelöydettävyys*. Tämä haaste tapahtuu muutamien muiden haasteiden seurauksena, joita sivuttiin tehdyissä haastatteluissa. Haaste *N. Metatietojen semanttiset ongelmat ohjelmissa* on yksi aiheuttavista haasteista. Tällä haasteella tarkoitetaan ohjelmien teknisen toiminnan olevan sellaista, että se ei ota aineistojen semanttisia asioita huomioon. Näin ollen esimerkiksi hakukoneiden indeksointi ja hakukoneoptimointi on hyvä ottaa huomioon avoimessa tiedossa.

”Aineistoissa kaikki tekstimäinen metatieto on Googlen indeksoimaa. Esimerkiksi tietokantataulukoiden kenttien arvot ovat Googlen löydettävissä ja tulee vastaan tietohauissa aineiston avoimuuden ansiosta. Nimenomaan metatiedot voivat tarttua hakuihin.” (B3)

Tätä ohjelmissa tapahtuvaa semanttisen tiedon käsittelyn vaikutusta aineistojen hakukonelöydettävyyteen kuvataan nuoliviivalla N-H.



Kuva 4: Syy-seuraussuhteet aineiston heikkoon hakukonelöydettävyyteen

Halutessaan hakukoneiden indeksointia ja näin metatietojen semanttisuuden näkymistä hakukonepalveluissa voidaan rajoittaa. Tämä kuitenkin heikentää aineiston löydettävyyttä. Aineistojen ja niitä hyödyntävien rajapintojen sekä ohjelmien käyttämät tiedostomuodot voivat olla sellaisia, että niiden sisältämät tekstimäiset kuvailut ovat näkyvissä. Nämä tulevat usein näkyviin hakukoneiden tuloksissa.

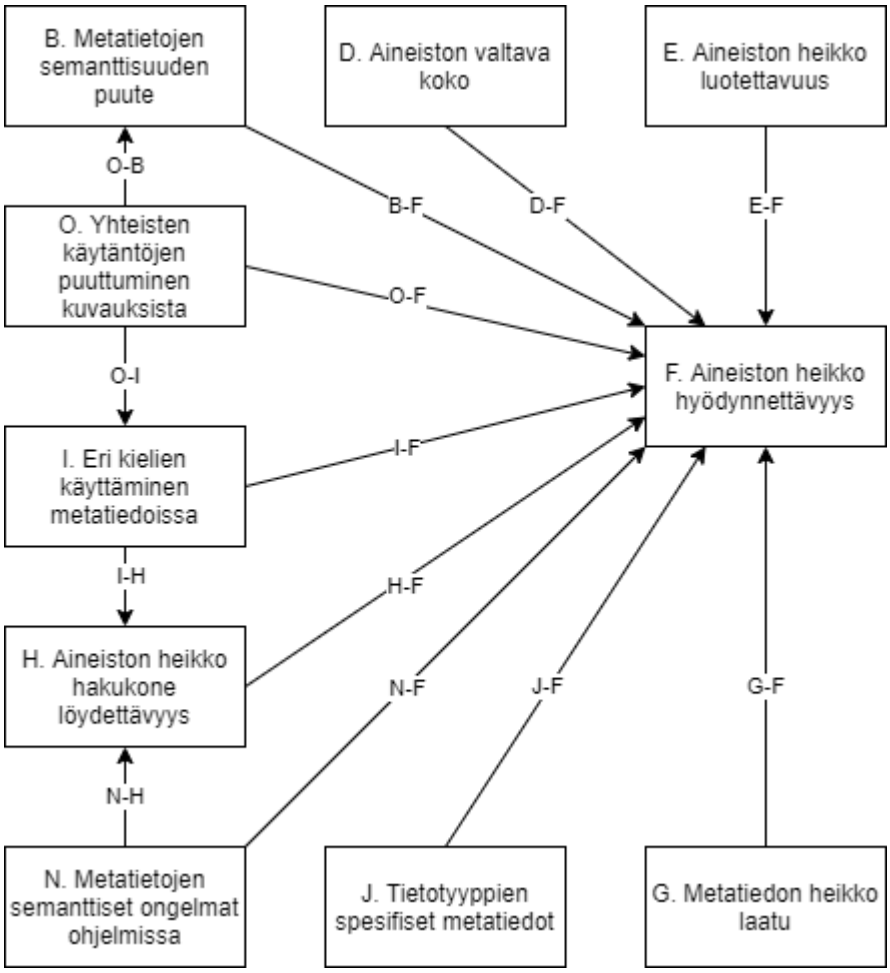
”Huono metatieto heikentää tiedon käyttöä. Jos vaikka omissa aineistoissa olisi estetty indeksointi ja ne eivät näkyisi Googlessa, niin käyttäjän olisi todella paljon vaikeampi löytää aineistoja, sillä ainoa keino löytää ne, olisi etsiä suoraan julkaisijan palvelusta.”
(B3)

Hakukoneiden tuloksiin vaikuttaa vahvasti aineistossa oleva sanasto ja kieli. Tätä yhteyttä kuvataan kuvan nuoliviivalla I-H. Hakukoneiden tuloksien lisäksi haaste *I. Eri kielen käyttäminen metatiedoissa* voi aiheuttaa itsessään ongelmia avoimelle tiedolle. On tilanteita, joissa aineiston tietorakenne, kuvaukset ja sisältö ovat eri kielillä. Tämä ei aiheuta suoraan ongelmia avoimelle tiedolle, mutta sen voidaan ajatella aiheuttavan epäjohdonmukaisuutta metatiedoissa.

Näitä sanastojen ja tietorakenteiden kielivalintoja voi olla määrittämässä eri standardit ja viitekehykset, joita käsiteltiin haasteessa P. Tätä yhteyttä kuvataan nuoliviivalla P-I.

4.1.4 Aineiston heikko hyödynnettävyys

Aineiston hyödynnettävyydellä tarkoitetaan sitä, onko avattu tieto käytettävää, löydettävää ja hyödyllistä sen käyttäjille. Tätä kokonaisuutta kuvaa haaste *F. Aineiston heikko hyödynnettävyys*, joka tapahtuu useiden eri muiden haastatteluissa esiin tulleiden haasteiden seurauksena. Nämä kuvataan seuraavassa kuvassa. Tämän haasteen voidaan nähdä olevan merkittävin kokoava haaste, johon muut haasteet vaikuttavat joko suoraan tai välillisesti. Edellä esiteltyjä haasteita ja yhteyksiä ei käsitellä uudelleen tämän verkon kuvauksessa selkeyden vuoksi.



Kuva 5: Syy-seuraussuhteet aineiston heikkoon käytettävyyteen

Edellä käsitelty haaste *B. Metatietojen semanttisuuden puute* vaikuttaa suoraan aineiston käytettävyyteen. Tämä ilmenee aineiston sisällön ja sen käyttötarkoituksen virheellisenä tulkintana, jota kuvataan nuoliviivalla B-F.

”On hyvin tärkeää tunnistaa, että rajapinnoissa on aina kaksi puolta, tekninen puoli ja sisällöllinen, eli semanttinen puoli. Ne ovat kaksi täysin eri asiaa. Usein ohjelmoijat haluavat hyödyntää jotain dataa nopeasti, eikä välttämättä olla edes niin kiinnostuttu sen tarkasta sisällöstä. Tällöin joutuu muistuttamaan, ettei se onnistu, jos ei sisältöä selvitä kunnolla. Semanttinen taso on todella tärkeä. Tämä on joskus unohtunut myös tiedon avaajilta globaalisti.” (B2)

Etenkin tilastollisessa datassa käyttäjillä tulisi olla tilastolliset perusmenetelmät hallinnassa. Muussa tapauksessa avattavan aineiston kuvauksissa ja ohjeistuksissa tulisi

käydä ilmi, miten aineistoa tulee käyttää. Tämä asettuu pääosin tiedon avaajan vastuulle. Myös aineiston vertailukelpoisuus on tärkeä saada aineiston hyödyntäjien tietoisuuteen.

Haaste *D. Aineiston valtava koko* vaikuttaa aineiston käytettävyyteen käytettyjen rajapintojen ja olemassa olevien dataportaalien kautta. Suuria aineistoja on haastava jakaa yleisten yksinkertaisten dataportaalien kautta esimerkiksi CSV-tiedostona, sillä niiden luominen ja käyttäminen on hyvin epätehokasta. Tämä tapahtuu etenkin paikkatiedon parissa, joka on itsessään muutenkin haastava jakaa yleisten dataportaalien kanssa, jota kuvataan tarkemmin haasteen J määrittelyssä. Aineiston suuri koko aiheuttaa haasteita teknisellä puolella yleisesti sen käsittelyssä ja siirtämisessä. Usein dataa otetaan kokonaisuudesta irti osissa, jolloin on vaarana, että aineiston metatiedon laatu heikkenee, tai niiden sisältämä automatiikka tai linkit katoavat. Tätä käsitellään myöhemmin lisää G-F syyseuraus-suhteen määrittämisessä.

Edellä käsitelty kokonaisuus haasteesta *E. Aineiston heikko luotettavuus* vaikeuttaa aineiston käytettävyyttä selvästi. Perimmäinen syy aineiston heikkoon luotettavuuteen on se, että tiedon käyttäjä ei tiedä mitä data on, mistä se on tullut ja mitä sillä voi tehdä.

”Tyypillisesti yleensä sillä, joka dataa etsii, on joku ongelma, jota hän yrittää ratkaista etsiessään aineistoja. Ne pitäisi kuvailla niin, että etsijä hahmottaa, että ratkaiseeko tämä ongelman vai ei.” (A2)

Tätä voidaan ehkäistä tekstimäisen kuvailun lisäksi visuaalisesti esimerkiksi näyttämällä suoraan aineiston sisältöä dataportaalissa. Mitä enemmän tietoa aineistosta käyttäjä saa helposti, sen paremmin hänelle tulee selväksi mistä data on peräisin, ja voiko kyseinen aineisto ratkaista hänen selvitettävän ongelman vai ei. Näin ollen aineiston heikko luotettavuus on suorassa yhteydessä heikon käytettävyyden kanssa, jota kuvataan nuoliviivalla E-F edellisessä kuvassa.

Aineiston metatiedon heikko laatu, eli haaste G, joka käsiteltiin edellä, aiheuttaa aineiston hyödynnettävyyteen ongelmia, jotka on kuvattu nuoliviivalla G-F. Metatiedon laatu voi olla alusta asti heikko, tai se voi heiketä esimerkiksi avoimen tiedon aineistosta otettujen otosten kautta. Tällöin metatiedon laatu on heikentynyt, jos se ei enää kuvaa aineistoa oikein ja täydellisesti. Esimerkiksi, jos jokin selittävä tekstimäinen kuvaus häviää otoksen mukana, tai poistuu kadonneen linkin seurauksena, on metatiedon täydellisyys mahdollisesti heikennyt. Vastaavasti, jos metatietoa ei päivitetä ja se vanhenee sellaisella tavalla, että se ei enää kuvaa aineistoa oikein, on metatiedon oikeellisuus kärsinyt.

Molemmissa tapauksissa aineiston käytettävyys on heikentynyt, sillä aineiston ymmärtämistä auttavat metatiedot eivät ole tarpeellisella tasolla.

Haaste *J. Tietotyypin spesifiset metatiedot* kuvaa avoimen tiedon ongelmia, jotka voivat aiheutua, kun aineiston sisältämä tieto on jotain hyvin spesifistä tietoa, jonka tietorakenne ja kuvailut sisältävät kyseiselle tietotyypille ominaisia erikoispiirteitä tai sanastoa. Esimerkiksi paikkatietoja sisältävät aineistot ovat hyvin spesifisiä johtuen niiden erityisistä aika- ja paikkadimensioista. Haaste aiheuttaa ongelmia sellaisille käyttäjille, joille kyseinen tietotyyppi ei ole ennestään tuttu. Avoimen tiedon aineistojen hyödyntäjä voi olla kuka tahansa, joten näitä ongelmia varmasti tapahtuu ajoittain. Näissä tapauksissa aineiston ymmärrettävyys ja sitä kautta hyödynnettävyys voi kärsiä. Tätä yhteyttä kuvataan nuoliviivalla J-F.

Avoimen tiedon yksi yleisimmistä hyödyntämisen käyttötapauksista on se, että yksi käyttäjä haluaa selvittää jonkin yhden spesifisen suureen tietomassasta tietyillä ehdoilla. Tietokoneohjelmat kuitenkin lukevat dataa ja koodia semanttisen tiedon sijasta, joten tämän kaltaisen tehtävän suorittaminen voi olla haasteellista, jos jokin avoimen tiedon tietorakenne ymmärretään vähänkin eri tavalla, tai mikäli jokin sanasto tulkitaan väärin. Näissä tapauksissa suuresta aineistosta pienen yksinkertaisen haun tulos voi osoittautua hankalaksi. Tämänkaltaisia ongelmia tiedon hyödyntämisessä kuvataan haasteella *M. Metatietojen semanttiset ongelmat ohjelmissa* ja kuvan nuoliviivalla N-F.

”Käyttöliittymän määrittelee sen sisältö. Eri sisällöille on eri käyttöliittymiä. Sisällöt voidaan jakaa useaan eri käyttöliittymään eri ryhmille ja käyttötarkoitukselle.” (B3)

Aineistolle voidaan luoda eri käyttötarkoituksiin omia palveluita tai jakelukanavia, jotka voivat helpottaa käyttäjien hakuja aineistoissa. Eri aineistoja voidaan käyttää usealla eri tavalla, mutta niiden käytössäkin voi piiletä omia haasteita.

”Jakelukanavia on paljon erilaisia. Avoimen datan käytettävyyden osalta on keskeistä itse palvelun kuvaaminen. Itse palveluun liittyvä metadata on haaste, sillä tietty rajapintapalvelu ja miten aineistoa pääsee hyödyntämään saattaa poiketa eri jakelutavoista hyvin paljon. Eli joudutaan kuvaamaan itse aineistoa, mutta myös palveluita, joiden kautta pääsee käsiksi aineistoihin.” (A1)

Seuraava käsiteltävä syy-seuraussuhde aineiston hyödynnettävyyteen on hakukonelöydettävyyden yhteys H-F. Tämä yhteys johtuu suoraan hakukoneiden mahdollisuudesta auttaa aineistojen löydettävyyttä ja sitä kautta hyödynnettävyyttä.

”Jos käyttäjät eivät löydä aineistoa, arvoa ei synny lainkaan.” (A1)

Hakukoneiden indeksoimat tiedot aineiston sisällöstä tai metatiedoista voivat näkyä suoraan hakujen tuloksissa. Tämä on merkittävä tekijä nykymaailmassa, jossa aineistoja varmasti haetaan yksittäisten dataportaalien ulkopuoleltakin. Hakukoneiden hakutuloksia auttaa suoraan jo se, kun tietoa avataan.

”Tiedon löydettävyyttä auttaa suoraan tiedon avaaminen. Jos tieto on suljettua, sitä ei näy tietenkään hakukoneissa. Asiakkuudet ja maksullisuus on myös merkittävä tekijä. Jos tieto ei ole suoraan avoimena, niin ei välttämättä tiedetä, onko tietoa olemassakaan.” (B3)

Aineistojen löydettävyys ei ole kuitenkaan pelkästään hakukoneista kiinni. Löydettävyyteen vaikuttaa paljon eri tekijöitä, joista osa on metatiedoista johtuvia ja osa ei. Metatiedot voivat olla löydettävyyden esteenä, jos ne eivät vastaa tarpeita.

”Tosi asiassa tietojen löytymiseen liittyy paljon parannettavaa siihen, että miten metatieto on olemassa, paljonko sitä on ja kuinka hyvää se on niin tietokannoissa kuin kuvauksissa. Mitä parempaa se metatieto on sitä paremmin ja tarkemmin ja enemmän tieto on käyttäjien löydettävissä.” (B3)

Löydettävyyden lisäksi myös aineiston ymmärrettävyys on merkittävä aineiston hyödyntämisen edellytys. Eri kielien käyttäminen voi heikentää aineiston ymmärrettävyyttä, esimerkiksi jos suomen kieltä puhumaton henkilö yrittää tulkita suomalaista dataa, tai jos suomalainen pyrkii tulkitsemaan ulkomaalaista dataa, jota ei ole kuvattu englannin kielellä. Tätä kuvaa haaste *I. Eri kielien käyttäminen metatiedoissa* ja sen syyseurausuhde I-F.

Viimeinen verkon käsiteltävä suhde on O-F, joka kuvaa haasteen *O. Yhteisten käytäntöjen puuttuminen metatiedoista* vaikuttavuutta aineiston hyödynnettävyyteen. Avoimen

tiedon standardit ja viitekehykset asettavat selkeyttäviä käytännön sääntöjä ja tietorakenteita, jotka auttavat tiedontarjoajia avaamaan tietoa. Kuitenkin tästä hyötyvät myös tiedon käyttäjät. Esimerkiksi yleisten lisenssien hyödyntäminen avatuissa aineistoissa helpottaa tiedontarjoajan taakkaa laatia käyttöehtoja, joita käyttäjien tulisi tulkita aineiston hyödyntämisessä.

”Vielä silloin kun ei noudatettu yleisiä viitekehyksiä ja lisenssejä pystyttiin tietoa avaamaan, mutta lisenssien käyttäminen helpottaa sitä suuresti avaajan lisäksi loppukäyttäjien näkökulmasta. Selkeät standardit helpottavat käyttöä.” (B3)

Tiedon avaamiseen vaikuttavat yleiset käytännöt näkyvät myös hallinnossa. Eri lait ja säädökset voivat suoraan ohjeistaa tiedon avaamista ja hyödyntämistä. Esimerkiksi PSI-direktiivin yksi kantava periaate on tukea avoimen tiedon käyttäjien kykyä löytää aineistot sekä helpottaa niiden käyttöä ilman hankaloittavia rajoituksia (Janssen 2011, s. 448).

”Ehdottomasti pitää mainita Euroopan Unionin PSI-direktiivi, jota on päivitetty monta kertaa. Sen periaatteet ovat vaikuttanut suoraan Suomeen ja muihinkin EU-maihin, etenkin niihin EU-maihin, joissa tietojen jakelu ei ole ollut lähtökohtaisesti avointa, kuten meillä Suomessa tänään on. Direktiivin vaikutus on ollut vuosien mittaan hyvin merkittävä ja aiheuttanut painetta vanhojen tietojen avaamiseen.” (B3)

Avoimen tiedon käyttöön ja sen ohjaamiseen vaikuttaa vahvasti se mitä sääntöjä ja käytäntöjä noudatetaan, sekä mikä lainsäädäntö on milloinkin voimassa.

4.2 Haastattelujen tulosten yhteenveto ja arviointi

Haastattelujen tuloksista nähdään, että metatiedoilla on runsaasti erilaisia vaikutuksia avoimeen tietoon. Haastatteluissa esiin nousseissa metatiedon haasteissa ja huomioon otettavissa asioissa oli paljon toisiinsa kohdistuvia syy-seuraussuhteita. Haasteet muodostavat näiden suhteiden avulla esitetyn verkon, jonka avulla voidaan hahmottaa, miten metatiedot vaikuttavat aineiston luotettavuuteen, hyödynnettävyyteen, hakukonelöydettävyyteen ja metatietojen laatuun.

Haastatteluissa oli yllättävää se, kuinka moni esiin nostettu haaste liittyi metatiedon semanttisuuteen. Aineistoja kuvailevat tekstimäiset metatiedot ovat siis selvästi hyvin merkityksellisiä avoimessa tiedossa. Avoimen tiedon haasteiden ja huomioon otettavien asioiden verkko sisältää yhden kokoavan haasteen, johon muut haasteet kohdistuvat suoraan tai välillisesti. Tämä haaste on *F. Aineiston heikko hyödynnettävyys*. Tätä haastetta tulisi pitää erityisen paljon silmällä avoimessa tiedossa. Muita merkittäviä haasteita olivat *E. Aineiston heikko luotettavuus*, *H. Aineiston heikko hakukonelöydettävyys* ja *G. Metatiedon heikko laatu*.

Haastattelujen tulosten arvioinnissa on hyvä pitää mielessä tutkimuksen aihepiirin monimutkaisuus. Avoin tieto ja niiden metatiedot eivät ole yleisessä tietoisuudessa. Tällä tarkoitetaan sitä, että haastatteluissa käytetty termistö, vastausten kohteet ja niiden yksityiskohdat ovat asioita, jotka ovat tuttuja vain ammattilaisille. Tästä syystä on mahdollista, että haastatteluissa on tapahtunut jotain väärinymmärryksiä. Esimerkiksi metatietotermin tarkka määrittely on hankalaa avoimessa tiedossa, joten tulosten käsittelyssä on voitu puhua eri asioista samoilla termeillä. Suoritetuissa haastatteluissa ei rajattu mitään metatietoon liittyvää huomiota pois.

Haastattelujen kriittisessä tarkastelussa voidaan todeta, että löydetyt avoimen tiedon metatietojen haasteet ja huomioon otettavat asiat kuuluvat varmasti tutkimuksen aihepiiriin, ja ne kaikki on hyvä ottaa tarkasteluun avoimen tiedon käsittelyssä. Kuitenkin niiden merkittävyyttä on vaikea arvioida. Lisäksi on tärkeää huomioida, että näitä haasteita ja huomioon otettavia asioita voi olla paljon lisääkin. Haastattelujen tulokset on tehty tiedon avaajien näkökulmasta, mikä tarkoittaa sitä, että tiedon hyödyntäjillä voi olla erilaisia huomioita.

5. POHDINTA

Tässä luvussa käsitellään tarkemmin haastattelujen tuloksia vertailemalla niitä teoriaan. Lisäksi pyritään tunnistamaan, onko samoja haasteita tunnistettu muissa tutkimuksissa. Avoimen tiedon metatietojen haasteita ei ole yleisellä tasolla listattu, joten pohdintaa tehdään ennalta tunnistettujen avoimen tiedon yleisten haasteiden ja metatietojen kriteereiden sekä ohjeistuksen perusteella. Tunnistettuihin avoimen tiedon haasteisiin ei esitetä ratkaisuja tässä tutkimuksessa, vaan niiden syitä ja vaikutuksia.

5.1 Aineistojen metatietojen laatu ja automaattisuus

Haastattelujen vastauksissa ja niiden avulla tunnistetuissa haasteissa nousi metatietojen laatu usein esiin eri asiayhteyksissä. Heikon metatiedon laadun voidaan nähdä vaikuttavan lähes kaikkiin haasteisiin, vaikka se on nostettu omaksi huomioksi tuloksissa. Metatiedon laadun tuleminen esiin vastauksissa ei ole yllättävää, ja sillä on oletettavastikin oma rooli aineiston hyödynnettävyydessä. Yksi haaste, joka on vahvasti linkittynyt heikkoon metatiedon laatuun, oli aineiston metatiedon automatiikan puute. Tämä on nostettu esiin myös tässä pohdinnan osuudessa, sillä se on huomioitu vahvasti myös muissa tutkimuksissa (Kubler *et al.* 2018) ja se tuli haastatteluissa esiin hyvin useasti.

Tämän tutkimuksen teoriaosuudessa mainituilla metatiedon laatutekijöillä ja teemahaastattelujen kautta saaduilla vastauksilla on yhteneväisyyksiä. Metatiedon laatutekijät täydellisyyden, oikeellisuuden ja relevanttiuden (Greenberg *et al.* 2008) suhteen tuli osassa haastatteluita esiin. Etenkin täydellisyyteen liittyviä ongelmia esiintyi vastauksissa. Esimerkiksi haaste *K. Metatietojen epäsopiva määrä* sisältää ongelmia, joissa metatiedot kokonaisuutena eivät kuvaa tarpeeksi kokonaisvaltaisesti aineistoa. Myös haaste *G. Metatiedon heikko laatu* ja siihen vaikuttavat tekijät liittyvät luonnollisesti metatietojen laatutekijöihin.

Metatietojen relevanttius ei suoraan noussut esiin haastatteluissa, mutta sen voidaan ajatella koskevan haastattelujen huomioita, jotka liittyvät aineistojen käyttäjien mahdollisuuteen hyödyntää avointa tietoa. Relevanttiuden voidaan nähdä vaikuttavan subjektiivisesti siihen, minkälainen metatieto on millekin käyttäjälle hyödyllistä. Suhteen N-F kuvailussa käytiin myös läpi, kuinka aineistoja voidaan jakaa eri jakelukanavia pitkin, ja näissä jakelukanavapalveluissa voi olla omat kuvailut ja metatiedot olemassa. Nämä metatiedot olisi hyvä luoda käyttäjälähtöisesti.

Metatietojen oikeellisuus kävi ilmi haastattelujen ajantasaisuuden ja pysyvyyden haasteissa L ja M. Haastattelujen perusteella oikeellisuus kärsii, jos metatietojen tiedot eivät enää pidä paikkaansa aineiston muuttuessa, kasvaessa tai siirtyessään jonkin rajapinnan läpi. Tätä voidaan ehkäistä hyödyntämällä automatiikkaa metatietojen päivittämisessä.

Metatietojen automaattisuus on tunnistettu yleisestikin tärkeäksi tekijäksi avoimessa tiedossa (Publications Office of the European Union. *et al.* 2020). Etenkin aineistojen koon voimakkaassa kasvussa metatietojen laaduntarkkailu ja automatiikka nousevat yhä tärkeämmiksi tekijöiksi aineiston hyödynnettävyydessä. Metatietojen ajan tasalla pysyminen sekä oikeellisuus eri rajapintojen puolilla on merkittävä tekijä aineistojen hyödynnettävyydessä.

Automaattisuutta on tutkittu myös muista näkökulmista. RDF-viitekehystä ja DCAT-sanastoa käyttävien aineistojen metatietojen laadun tarkkailun automaattisuutta on tutkittu mm. avoimen tiedon portaaleissa (Kubler *et al.* 2018). Vaikka tämän tutkimuksen haastattelut suoritettiin yleisellä tasolla, tulivat avoimen tiedon portaalit osissa vastauksissa esiin. Teoriaosuudessa mainitut viisi tarkkailtavaa metatiedon sisällön kriteeriä antavat tietoa siitä, millaisia seikkoja avoimen tiedon portaaleissa on hyvä huomioida metatietojen kannalta. Nämä kriteerit ovat olemassaolo, vaatimustenmukaisuus, saatavuus, tarkkuus sekä avoin tieto (Neumaier et al. 2016).

Olemassaolo -kriteerillä tarkoitettiin sitä, että sisältääkö metatieto kaikki tarvittavat kentät. Tämä viittaa vahvasti metatietojen täydellisyyteen, joka käsiteltiin jo edellä. Lisäksi *tarkkuus* -kriteerin voidaan ajatella liittyvän suoraan edellä mainittuun metatiedon oikeellisuuteen. *Vaatimustenmukaisuus* -kriteeri viittaa siihen, käyttääkö ja noudattaako metatieto sille asetettuja vaatimuksia. Tämä tuli esiin tunnistetussa haasteessa *O. Yhteisten käytäntöjen puuttuminen metatiedoista*. Haastatteluissa kävi ilmi, että standardien ja viitekehysten puuttuminen aiheuttaa paljon ongelmia aineiston hyödynnettävyyteen. Standardien ja viitekehysten noudattamisesta nousi esiin se, että se on työlästä, mutta silti arvokasta.

Saatavuus -kriteeriin liittyviä haasteita ei noussut haastatteluissa pinnalle. Suuri osa haasteista liittyi metatietojen sisältöön, eikä niinkään siihen, että miten ne ovat saatavilla. Tämä ei haastattelujen perusteella vaikuta suoraan aineiston hyödynnettävyyteen, mutta se on hyvä piirre avoimen tiedon metatiedoissa. Tämän voidaan nähdä olevan haaste, joka ei haastatteluissa käynyt ilmi.

Viimeisenä oleva *avoin tieto* -kriteeri kuvaa metatiedon ominaisuuksien olevan juuri avoimelle tiedolle sopivia. Haastattelut suoritettiin avoimen tiedon näkökulmasta, joten kaikkien vastausten voidaan nähdä koskettavan tätä kriteeriä. Kuitenkaan mikään vastaus tai tunnistettu haaste ei suoraan osoittanut tähän kriteeriin. Voidaan kuitenkin pohtia sellaista tilannetta, jossa jotain vanhaa olemassa olevaa aineistoa halutaan jälkikäteen avata, jolloin on pohdittava sen metatietoja. Tällaisia tilanteita on varmasti ollut haastateltavilla organisaatioilla, mutta niitä ei käsitelty tarkemmin haastattelujen kysymysten tai vastausten puitteissa.

Metatiedon laadun varmistaminen ja ylläpitäminen on vahva tekijä aineiston hyödynnettävyydessä. Automatiikan avulla voidaan varmistaa, että aineiston metatieto kuvaa oikein aineiston sisältöä. Haastattelujen perusteella metatiedon heikkoon laatuun on linkittynyt metatietojen ajantasaisuuden puute, pysyvyyden puute sekä metatietojen epäsopiva määrä. Näihin keskittymällä on mahdollista kehittää metatiedon laatua ja sitä kautta aineiston hyödynnettävyyttä. Metatietojen laadun automatiikalla pitäisi keskittyä juuri metatietojen ajantasaisuuden ja pysyvyyden ylläpitämiseen sellaisilla tietoteknisillä menetelmillä, kuten esimerkiksi Kubler et. al on tutkinut (2018).

5.2 Aineistojen löydettävyys ja hakukoneoptimointi

Haastattelujen vastauksissa kävi ilmi, kuinka tärkeää on antaa käyttäjille mahdollisuus löytää avatut aineistot. Ilman aineiston löytämistä ja hyödyntämistä, niiden arvo ei tule käyttöön. Hakukoneilla on suuri rooli parantaa aineistojen löydettävyyttä etenkin sellaisissa tilanteissa, jossa aineistoja ei etsitä aineistoille spesifisten avoimen tiedon portaalien kautta.

Yleisellä tasolla hakukonelöydettävyys on tunnistettu olevan merkittävä tekijä avoimen tiedon hyödynnettävyydessä (Publications Office of the European Union. *et al.* 2020, s. 39). Aineiston potentiaalinen arvo tulee sitä paremmin hyödynnettyä, mitä enemmän sitä käytetään. Käyttäminen vaatii aineiston löytämistä, johon laadukkaalla ja suunnitellulla metatiedoilla voidaan vaikuttaa tämän tutkimuksen haastattelujen perusteella.

Perinteisten hakukoneiden lisäksi on tehty erityisesti avoimen tiedon hakemiseen suunniteltuja hakukoneita. Näistä yksi esimerkki on Google Dataset Search. Tämän kaltaisten hakukoneiden hakutulosten kehittäminen metatietojen hakukoneoptimoinnilla voi parantaa aineiston hyödynnettävyyttä.

5.3 Aineistojen luotettavuus

Julkisten avoimien aineistojen luotettavuus ja läpinäkyvyys on vahvasti esiin tullut haaste tässä tutkimuksessa. Sen voi ajatella olevan myös yhtenä taustatekijänä *Tiedon hyödyntäminen ja avaaminen* –hankkeessa, ja se mainitaan myös PSI-direktiivissä (EU 2019/1024). Aineiston metatietojen vaikuttavuus aineiston luotettavuuteen on vaikea arvioida tarkasti, mutta sillä on selvästi vaikutusta.

Myös teoriaosuudessa mainitussa metatietojen merkittävien periaatteiden listassa Tauberer (2014b) mainitsee aineiston luotettavuuden olevan tärkeä tekijä. Tiedon tulee olla läpinäkyvästi julkaistuna, eikä sitä pidä kyetä jälkikäteen muokkaamaan ilman muokkauksesta jäävää jälkeä. Haastatteluissa olevat aineiston luotettavuuteen vaikuttavat haasteet liittyivät usein käyttäjien käsitykseen siitä, mihin aineistoa voi hyödyntää, ja mihin ei. Se, mikä tekee aineistosta tarkalleen luotettavan, on subjektiivinen käsitys ja vaatii erillistä tarkempaa tutkimusta.

5.4 Aineistojen semanttisuus ja yhdistetty avoin tieto

Teorialuvussa esitettyä yhdistettyä avointa tietoa (engl. linked open data) voidaan pitää yhtenä avoimien aineistojen kehityksen päämääristä. Sen avulla on mahdollista yhdistää aineistoja globaalilla tasolla, ja luoda suoria datayhteyksiä erilaisten toimialueiden (domain) välille. (Sikos 2015) Tämän tutkimuksen haastatteluissa esiin nousseiden semanttisuuteen liittyvien haasteiden määrä antaa viitteitä siitä, että avoimen tiedon metatietojen semanttisuuteen on kiinnitetty paljon huomiota Suomessa. Esimerkiksi tunnistetuista haasteista *B. Metatietojen semanttisuuden puute* ja kaikki siihen yhteydessä olevat haasteet kuuluvat tähän aihepiiriin, ja vaikuttavat selvästi aineiston luotettavuuteen ja hyödynnettävyyteen.

Yhdistetty avoin tieto liittyy pääosin aineistojen tekniseen ja semanttiseen yhteentoimivuuteen. Tavoitteena on luoda tilanne, jossa kaikki avoimet aineistot olisivat teknisesti toisissaan kytköksissä globaalisti niin, että niitä voisi käsitellä ja hyödyntää yhdessä (Sikos 2015), mahdollisesti pilvipalveluissa (Gandon *et al.* 2015). Haastattelun tuloksissa olevat semanttiset haasteet ja huomioon otettavat asiat koskivat usein käyttäjien näkökulmaa sekä aineistojen ymmärrettävyyttä. Samat asiat lienevät yhtä relevantteja yhdistetyssä avoimessa datassa, jossa metatietojen semanttisuus on edelleen se tekijä, joka vaikuttaa vahvasti aineiston ymmärrettävyyteen ja luotettavuuteen. Käytettävien aineistojen määrän kasvaessa voisi ajatella, että semanttisuuden merkitys voi kasvaa vieläkin suuremmaksi.

Haastatteluissa ei suoraan esiintynyt mainintoja metatietojen teknisestä ja semanttisesta yhteentoimivuudesta, vaan se ilmeni välillisesti esiin tulleista huomioista. Yhdistetty avoin tieto tai semanttinen web mainittiin haastattelujen aikana muutaman kerran, mutta sitä aihetta ei yhdistetty mihinkään haasteeseen tai huomioon otettavaan asiaan missään vaiheessa. Tästä voidaan päätellä, että suurin osa pinnalla olevista metatietojen haasteista liittyy organisaatioiden omien aineistojen kehittämiseen ja ylläpitoon heidän käyttäjiensä näkökulmasta. *Tiedon hyödyntäminen ja avaaminen* -hankkeessa yhtenä tavoitteena on kehittää avattavan tiedon laatua ja yhteentoimivuutta teknisesti ja semanttisesti (Valtiovarainministeriö 2020). Yhdistetty avoin tieto ei varsinaisesti ole osa hankkeen tavoitteita, vaan se esiintyy yleisessä aihepiirin teoriassa sekä muissa tutkimuksissa.

Tästä voidaan päätellä, että tällä hetkellä kansallisesti keskitytään omien aineistojen kehittämiseen ja ylläpitoon, sekä uusien aineistojen avaamiseen. Lisäksi ollaan luomassa linjauksia julkisen tiedon avaamiselle ohjelmointirajapintojen kautta. Samalla pyritään myös ottamaan käyttöön avoimen tiedon laatukriteerit (Valtiovarainministeriö 2020). Yhdistettyä avointa tietoa voidaan pitää seuraavana vaiheena, jossa aineistojen hyödyllisyyttä pohditaan globaalilla tasolla.

Teoriaosuudessa NISO:n mainitsemat kuusi avoimen tiedon avaamisen metatiedon periaatetta (2007) ottavat huomioon pitkän aikavälin suunnitelmallisuuden metatietojen käytössä. Tähän liittyy esimerkiksi metatietojen viitekehykset, standardit ja sanastot. On esitetty, että esimerkiksi RDF-viitekehystä käyttäessä mahdollistetaan yhdistetyn avoimen tiedon käyttäminen (Sikos 2015). Haastattelujen perusteella tiedetään, että tämän kaltaisia viitekehyksiä käytetään yleisesti jo Suomessa, mutta eivät kaikki ainakaan vielä globaalilla tasolla. Yhdistettyyn avoimeen tietoon on siis varauduttu suoraan tai välillisesti yleisten standardien ja käytäntöjen puitteissa.

Haastattelussa C1 mainitsee myös, että avoimen tiedon politiikan olemassaolon pitäisi vaikuttaa siihen, että avointa tietoa ja sen rajapintoja pitäisi käsitellä niin, että ne ovat alusta asti suunniteltu avoimiksi. Samalla periaatteella voidaan pohtia, pitäisikö Suomessa avoimen datan politiikassa ottaa alusta asti metatietojen globaalit tekijät huomioon. Toisaalta tämän tutkimuksen haastatteluissa ja tavoitteissa oli tähtäimenä selvittää avoimen tiedon metatietojen haasteita ja huomioon otettavia asioita yleisellä tasolla. Haastattelujen tuloksien perusteella ei voida arvioida yhdistetyn avoimen tiedon haasteita. Mainittakoon kuitenkin, että haastattelujen perusteella voidaan sanoa, että Suomessa käytetään merkittävän paljon yleisesti käytettyjä globaaleja standardeja, viitekehyksiä ja sanastoja avoimessa tiedossa.

Yhdistettyä avointa tietoa pitäisi kuitenkin tavoitella ja pitää tähtäimenä avoimen tiedon kehityksessä globaalilla tasolla. Haastatteluissa esiin tulleissa haasteissa merkittävä osa liittyi metatietojen semanttisuuteen. Haaste *B. Metatietojen semanttisuuden puute* ja siihen ketjuutuneet haasteet A, O, P ja Q ovat osasyinä aineiston heikkoon hyödynnettävyyteen ja yhdistetyn avoimen tiedon puutteellisuuteen. Niihin keskittymällä voidaan edistää yhdistetyn avoimen tiedon kattamista ja aineistojen pitkän aikavälin hyödynnettävyyttä.

6. YHTEENVETO JA PÄÄTELMÄT

Tässä laadullisessa tapaustutkimuksessa selvitettiin avoimen tiedon metatietojen haasteita ja huomioon otettavia asioita yhdistelemällä olemassa olevaa tieteellistä teoriataustaa sekä suoritettuja teemahaastatteluja. Haastattelut käsiteltiin induktiivisella sisällön analyysilla ja vastauksista tunnistettiin 17 eri haastetta ja huomiota. Näillä haasteilla oli selvät syy-seuraussuhteet, jotka kuvattiin havainnollisella verkolla.

Avoimen tiedon merkitys kasvaa yhä enemmän. *Tiedon hyödyntäminen ja avaaminen* -hankkeen ja PSI-direktiivin tapaiset tekijät kiihdyttävät kansallisen avoimen tiedon hyödynnettävyyttä entisestään. Avoimen tiedon hyödyntämisessä metatiedot ovat avain roolissa. Metatiedoissa on kuitenkin olemassa haasteita ja huomioon otettavia asioita, joita tässä tutkimuksessa on tarkoituksena selvittää.

Tämän tutkimuksen tutkimuskysymyksenä oli *”Mitä metatietojen haasteita ja huomioon otettavia asioita on olemassa avoimen tiedon avaamisessa ja hyödyntämisessä?”.* Tähän kysymykseen voidaan vastata haastattelujen tuloksilla ja tunnistetuilla haasteilla, jotka on kuvattu taulukossa 2. Näitä tarkennettiin kuvassa 1 havainnollistavassa haasteiden syy-seuraussuhteiden verkossa, jotka purettiin osiin ja käsiteltiin yksityiskohtaisesti.

Tunnistetut 17 haastetta ovat ketjuutuneet toisiinsa ja vaikuttavat välillisesti neljään kokoavaan haasteeseen, jotka ovat metatiedon heikko laatu, aineiston heikko luotettavuus, heikko hakukonelöydettävyys ja heikko hyödynnettävyys. Näillä haasteilla on suuri merkitys avoimen tiedon arvonluonnissa. Haasteiden syy-seuraussuhteiden kuvaavalla verkolla voidaan kuvata miten löydetyt haasteet vaikuttavat näihin kokoaviin haasteisiin. Haasteiden syitä ja vaikutuksia voidaan arvioida verkon avulla, sekä arvioida miten eri haasteiden seuraukset ketjuuntuvat eteenpäin lopulta heikentäen aineiston hyödynnettävyyttä.

Tämän tutkimuksen tuloksista voidaan nostaa kolme mielenkiintoista asiaa ylös:

- Avoimen tiedon metatietojen haasteet ovat ketjuutuneita ja toisiinsa yhteydessä.
- Haasteet kohdistuvat neljään merkittävään haasteeseen ja huomioon, jotka ovat aineiston heikko luotettavuus, heikko hakukonelöydettävyys, heikko hyödynnettävyys ja metatietojen heikko laatu.
- Avoimen tiedon haasteet ja huomioon otettavat asiat ovat useimmiten semanttisia – ei teknisiä.

Tunnistettujen haasteiden ja pohdinnan avulla tämä tutkimus on tuonut uutta tietoa metatietojen haasteista avoimessa tiedossa.

6.1 Tutkimuksen arviointi

Tutkimuksessa merkittävimmät havainnot syntyivät tutkimuksen haastatteluista. Näistä haastatteluista saatiin uutta tietoa. Haastateltavat henkilöt olivat hyvin asiantuntevia ja haastatteluissa esiin nousseet avoimen tiedon metatietojen haasteet ja huomioon otettavat asiat olivat hyvin arvokkaita. Haastatteluja tehtiin kolmen eri julkisen tietoa avaavan organisaation näkökulmasta. Haastattelumäärä jäi kohtalaisen vähäiseksi, ja suuremman haastatteluotannan kanssa metatiedoista olisi voitu saada vielä laajemmin tietoa.

Tapaustutkimusten validiteettia ja reliabiliteettia on vaikea arvioida. Tutkimuksen läpinäkyvyyttä pyrittiin parantamaan tuomalla esiin empiriaosuuden haastattelujen vastausten suoria lainauksia. Tutkimuksessa käytetty induktiivinen sisällön analysointi on tulkitsevaa ja subjektiivista toimintaa, joka voi aiheuttaa käsiteltyjen avoimen tiedon metatietojen haasteiden syiden ja vaikutusten väärinymmärryksiä.

6.2 Jatkotutkimuskohteet

Haastattelussa esiin nousseista asioista luotiin 17 verkkoon asetettua haastetta ja huomioon otettavaa asiaa. Tämän kaltaista verkkoa olisi hyvä esitellä toisille avoimen tiedon ja metatietojen asiantuntijoille tarkasteltavaksi ja tutkittavaksi. Olisi mielenkiintoista kuulla, tunnistavatko he samoja haasteita, vai eivät.

Kappaleessa 2.2 esitettiin kuusi arvokasta tietovarantotyyppiä. Haastateltavat organisaatot edustivat yhteensä neljää näistä. Arvokkaat tietovarantotyypit tulivat tämän seurauksena suurimmilta osin käsiteltyä, mutta puuttuvista kahdesta tyypistä (*yritys- ja yritysten omistustiedot* sekä *liikkuvuustiedot)* olisi voinut syntyä vielä uusia huomioita haastatteluissa. Mahdollisessa jatkotutkimuksessa nämä käsittelemättä jääneet tietovarantotyypit olisi hyvä ottaa mukaan.

Toinen jatkotutkimusaihe olisi haastatella samasta aiheesta tiedon hyödyntäjiä. Avoimen tiedon tarjoajilla ja niiden hyödyntäjillä voi olla paljon eri näkemyksiä millaista metatiedon tulee olla aineistoissa. Tiedon hyödyntäjillä ja muilla asiantuntijoilla voi olla myös arvokkaita ratkaisuehdotuksia kyseisiin haasteisiin.

LÄHTEET

Aroyo, L. et al. (2011) The Semantic Web – ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part II. doi: 10.1007/978-3-642-25093-4.

Attard, J. et al. (2015) A systematic review of open government data initiatives, Government Information Quarterly, 32(4), pp. 399–418. doi: 10.1016/j.giq.2015.07.006.

Attard, J., Orlandi, F. & Auer, S. (2016) Value Creation on Open Government Data. 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 2605–2614. doi: 10.1109/HICSS.2016.326.

Avoindata.fi (2021) Mitä on avoin data?, Avoindata.fi. Saatavissa: https://www.avoindata.fi/fi/opas/mita-on-avoin-data (Viitattu: 26. huhtikuuta 2021).

Cox, S. (2018) RDF representation of 2017 edition of International Chronostratigraphic Chart (Geologic Timescale). CSIRO. doi: 10.25919/5B4D2B83CBF2D.

Dey, I. (1993) Qualitive data analysis. A user-friendly guide for social scientist. Routledge, London.

European Comission (2019) Euroopan parlamentin ja neuvoston direktiivi (EU) 2019/ 1024, - annettu 20 päivänä kesäkuuta 2019, - avoimesta datasta ja julkisen sektorin hallussa olevien tietojen uudelleenkäytöstä, p. 28.

European Comission (2020) From the Public Sector Information (PSI) Directive to the open data Directive. Saatavissa: https://ec.europa.eu/digital-single-market/en/public-sector-information-psi-directive-open-data-directive (Viitattu: 14. huhtikuuta 2021)

European Comission (2021) About INSPIRE | INSPIRE. Saatavissa: https://inspire.ec.europa.eu/about-inspire/563 (Viitattu: 26. huhtikuuta 2021).

Finto (2020) Finto: YSO: yhdistetty avoin tieto. Saatavissa: https://finto.fi/yso/fi/page/p26001 (Viitattu: 26. huhtikuuta 2021).

Forsström, P.-L. (2019) Uudistunut PSI-direktiivi tuo uutta puhtia saatavuuteen, Avoin tiede. Saatavissa: https://avointiede.fi/fi/ajankohtaista/uudistunut-psi-direktiivi-tuo-uutta-puhtia-saatavuuteen (Viitattu: 6. lokakuuta 2020).

Gandon, F. et al. (2015) The Semantic Web: ESWC 2015 Satellite Events: ESWC 2015 Satellite Events. doi: 10.1007/978-3-319-25639-9.

Gilliland, A. J. (2008) Setting the Stage. Saatavissa: http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf (Viitattu: 1. huhtikuuta 2021).

Greenberg, J. et al. (2008) Metadata for semantic and social applications: proceedings of the International Conference on Dublin Core and Metadata Applications.

Janssen, K. (2011) The influence of the PSI directive on open government data: An overview of recent developments, Government Information Quarterly, 28(4), ss. 446–456. doi: 10.1016/j.giq.2011.01.004.

Kallinen, T. & Kinnunen, T. (2021) Tapaustutkimus, Tietoarkisto. Saatavissa: https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/tutkimusasetelma/tapaustutkimus/ (Viitattu: 8. toukokuuta 2021).

Kubler, S. et al. (2018) Comparison of metadata quality in open data portals using the Analytic Hierarchy Process, Government Information Quarterly, 35(1), ss. 13–29. doi: 10.1016/j.giq.2017.11.003.

Kyngäs, H. & Vanhanen, L. (1999) Sisällön analyysi. Oulun yliopisto.

National Information Standards Organization (2007) A Framework of Guidance for Building Good Digital Collections - 3rd edition, p. 100.

National Information Standards Organization (2021) What We Do | NISO website. Saatavissa: https://www.niso.org/what-we-do (Viitattu: 6. toukokuuta 2021).

Neumaier, S., Umbrich, J. & Polleres, A. (2016) Automated Quality Assessment of Metadata across Open Data Portals, Journal of Data and Information Quality, 8(1), ss. 1–29. doi: 10.1145/2964909.

Niininen, S., Nykyri, S. & Suominen, O. (2017) The future of metadata: open, linked, and multilingual – the YSO case, Journal of Documentation, 73(3), ss. 451–465. doi: 10.1108/JD-06-2016-0084.

Open Knowledge Fountain (2021) Avoimen tiedon määritelmä - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. Saatavissa: http://opendefinition.org/od/1.1/fi/ (Viitattu: 26. huhtikuuta 2021).

Pomerantz, J. (2015) Metadata. Cambridge, UNITED STATES: MIT Press. Saatavissa: http://ebookcentral.proquest.com/lib/tampere/detail.action?docID=4397948 (Viitattu: 9. marraskuuta 2020).

Publications Office of the European Union. et al. (2020) Reusing open data: a study on companies transforming open data into economic and societal value. LU: Publications

Office. Saatavissa: https://data.europa.eu/doi/10.2830/876679 (Viitattu: 9. toukokuuta 2021).

Riley, J. (2017) Understanding metadata: what is metadata, and what is it for? Saatavissa: http://www.niso.org/publications/understanding-metadata-riley (Viitattu: 9. marraskuuta 2020).

Sandelowski, M. (1993) Focus on qualitive methods. Qualitive analysis: What is it and how to begin. Research in Nursing & Health 18, ss. 371-375.

Saunders, M., Lewis, P. & Thornhill, A. (2019) Research Methods for Business Students Ebook. Saatavissa: http://ebookcentral.proquest.com/lib/tampere/detail.action?docID=5774742 (Viitattu: 8. toukokuuta 2021).

Sikos, L. F. (2015) Chapter 3: Linked Open Data - Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data.

Sugimoto, S. (2014) Digital archives and metadata as critical infrastructure to keep community memory safe for the future – lessons from Japanese activities, Archives and Manuscripts, 42(1), ss. 61–72. doi: 10.1080/01576895.2014.893833.

Suomidigi (2020) JHS 189 Avoimen tietoaineiston käyttölupa, Suomidigi. Saatavissa: https://www.suomidigi.fi/ohjeet-ja-tuki/jhs-suositukset/jhs-189-avoimen-tietoaineiston-kayttolupa (Viitattu: 4. toukokuuta 2021).

Tauberer, J. (2014a) 14 Principles of Open Government Data - Open Government Data: The Book. Saatavissa: https://opengovdata.io/2014/principles/ (Viitattu: 26. huhtikuuta 2021).

Tauberer, J. (2014b) Open Government Data: The Book. Saatavissa: https://opengovdata.io/ (Viitattu: 26. huhtikuuta 2021).

Thomas, D. R. (2003) A general inductive approach for qualitative data analysis, s. 11.

UNECE (2021) GSIM and standards - GSIM and standards - UNECE Statswiki. Saatavissa: https://statswiki.unece.org/display/gsim/GSIM+and+standards (Viitattu: 26. huhtikuuta 2021).

Valtiovarainministeriö (2020) Tiedon hyödyntämisen ja avaamisen hanke 2020-2022. Hankesuunnitelma. Valtiovarainministeriö.

Weber, R.P. (1985) Basic Content Analysis. Sage Publications, Newbury Park.

W3C (2014) RDF - Semantic Web Standards. Saatavissa: https://www.w3.org/RDF/ (Viitattu: 8 toukokuuta 2021).

W3C (2020) Data Catalog Vocabulary (DCAT) - Version 2. Saatavissa: https://www.w3.org/TR/vocab-dcat-2/ (Viitattu 26. huhtikuuta).

Zhang, A. B. and Gourley, D. (2008) Creating Digital Collections.

Zuiderwijk, A. et al. (2012) Socio-technical Impediments of Open Data, Electronic Journal of E-Government: EJEG, 10(2), ss. 156–172.

LIITE A: HAASTATTELUKYSYMYKSET

1) Peruskysymykset
    a) Organisaatio lyhyesti
    b) Haastateltavan rooli
    c) Avattavan tiedon tyyppi
    d) Avattu rajapinta lyhyesti
2) Mitä metatietojen haasteita organisaatiolla on tullut vastaan tiedon avaamisessa?
    a) Haasteen kuvaus
    b) Mitä vaiheita tiedon avaamiseen kuuluu ja missä vaiheessa haaste ilmeni?
    c) Haasteen syyt, seuraukset ja mahdolliset ratkaisut?
    d) Mitä on syytä huomioida kyseisessä tilanteessa?
    e) Onko haaste organisaatiosta tai avatun tiedon tyypistä riippuvainen?
3) Mitä metatietojen haasteita tiedon hyödyntämisessä on tullut vastaan?
    a) Haasteen kuvaus
    b) Mitä vaiheita tiedon avaamiseen kuuluu ja missä vaiheessa haaste ilmeni?
    c) Haasteen syyt, seuraukset ja mahdolliset ratkaisut?
    d) Mitä on syytä huomioida kyseisessä tilanteessa?
    e) Onko haaste organisaatiosta tai avatun tiedon tyypistä riippuvainen?