Oskar Lindholm

Automaattisten tiedonhankintamenetelmien soveltuvuus avointen lähteiden tiedustelussa

Tietotekniikan kandidaatintutkielma

19. toukokuuta 2023

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Oskar Lindholm

Yhteystiedot: oskar.g.lindholm@student.jyu.fi

Ohjaaja: Timo Tiihonen

Työn nimi: Automaattisten tiedonhankintamenetelmien soveltuvuus avointen lähteiden tiedustelussa

Title in English: Automated Intelligence Collection Methods in Open-source Intelligence

Työ: Kandidaatintutkielma

Opintosuunta: Tietotekniikka

Sivumäärä: 48+0

Tiivistelmä: Avointen lähteiden tiedustelun haasteeksi on muodostunut digitaalisen tiedon alati kasvava määrä, joka vaikeuttaa relevantin tiedon löytämistä internetistä. Tiedonhankintaa automatisoimalla tiedustelu kykenee keskittämään enemmän resursseja kerätyn tiedon käsittelyyn ja analysointiin. Tässä tutkielmassa tarkastellaan automaattisten tiedonhankintamenetelmien hyödyntämistä verkossa tapahtuvassa avointen lähteiden tiedustelussa. Tutkielman tavoitteena oli selvittää tiedonhankintamenetelmien soveltuvuutta erilaisiin tiedontarpeisiin. Teksti- ja kuvatulkintamenetelmät osoittautuivat tutkimuksen perusteella monikäyttöisimmiksi tiedonkeräysmenetelmiksi. Tutkimus toteutettiin kirjallisuuskatsauksena.

Avainsanat: avointen lähteiden tiedustelu, OSINT, automaatio, tiedonkeräysmenetelmät

Abstract: The rapidly increasing amount of digital information is becoming a major challenge for open-source intelligence practitioners in finding relevant information. By automating the tasks of intelligence collection, intelligence practitioners are able to allocate more resources into processing and analysing the collected information. This bachelor's thesis focuses on the use of automated intelligence gathering methods in web-based open-source intelligence. The purpose of the study was to evaluate the suitability of automated intelligence gathering methods to different types of information needs. Text- and image-based gathering methods turned out to be the most versatile methods in meeting different information

needs. The research was conducted as a literature review.

Keywords: open-source intelligence, OSINT, automation, intelligence collection methods

Termiluettelo

avoin lähde	Lähde, johon kenellä tahansa on mahdollisuus hankkia pääsy ilmaiseksi, pyynnön tai maksun kautta.
avointen lähteiden tiedustelu	Tiedustelulaji, jossa keskitytään hankkimaan tietoa avoimista lähteistä, kuten massamediasta, kirjallisuudesta tai julkisista tietokannoista.
informaatio	Informaatio on tietoa, jolla on konteksti ja rakenne, jotka antavat tiedolle merkityksen. Informaatiota voidaan tulkita erilaisista perspektiiveistä tietämyksen luomiseksi.
OSINT	Yleisesti käytetty lyhenne avointen lähteiden tiedustelua tarkoittavasta englanninkielisestä termistä *open-source intelligence*.
metatieto	Tietoa, joka kuvailee tiivistetysti toista tietosisältöä. Esimerkiksi tiedostoa kuvailevaa metatietoa ovat tiedoston laatijan nimi, tiedostopääte ja viimeisin tallennusajankohta.
tiedontarve	Kokemus toimintaympäristöön tai tilanteeseen liittyvästä epävarmuudesta, jota pyritään vähentämään hankkimalla tietoa ja parantamalla näin tilannetietoisuutta.
tiedustelu	Tiedustelu on päätöksentekoa tukevaa tiedonhankintaa, tiedon analysointia ja uuden tiedon tuottamista. Tiedustelun tarkoituksena on lisätä tai kehittää päätöksentekijän ymmärrystä omassta toimintaympäristöstään.
tieto	Tieto on klassisessa merkityksessään ”hyvin perusteltu, tosi uskomus”. Tieto voidaan jakaa sen jalostusasteen mukaan dataksi, informaatioksi, tietämykseksi ja ymmärrykseksi.

Kuviot

Sisällys

1 Johdanto

Modernin tietoyhteiskunnan tunnusomainen piirre on laaja-alainen tietojen kerääminen, jakaminen ja hyödyntäminen yhteiskunnan eri sektoreilla. Tiedon tehokas hyödyntäminen edellyttää kykyä tunnistaa ja toimittaa relevanttia tietoa päättäjille, jotka soveltavat sitä organisaatioidensa toiminnan johtamisessa (Martelius 2020). Tiedon jakamista ja päätöksentekoa tuetaan tiedustelutoiminnalla, jossa kerätään, analysoidaan ja koostetaan tietoa toimintaympäristön olosuhteista. Tiedustelua hyödyntäviä tahoja ovat esimerkiksi kansainväliset järjestöt ja valtioiden hallinnot (Best 2008), turvallisuusviranomaiset (Bayerl ym. 2023) sekä kaupalliset yritykset (Yang ja Lee 2012). Tietoja kerätään useista lähteistä erilaisilla tiedonhankintamenetelmillä kattavan ja tarkan kokonaiskuvan muodostamiseksi toimintaympäristöstä.

Avointen lähteiden tiedustelu on tiedustelulaji, jossa kerätään tietoja vapaasti saatavilla olevista lähteistä, kuten sanomalehdistä, tv-ohjelmista ja julkisista tietokannoista (Best 2008). Digitalisaation seurauksena valtaosa julkisesti saatavilla olevasta informaatiosta on nykyään löydettävissä internetistä. Internetin käytön yleistymisen ja lähdemateriaalin räjähdysmäisen kasvun myötä avointen lähteiden tiedustelusta on tullut aiempaa käytännöllisempi tiedustelulaji. Vapaasti saatavissa olevan tiedon suurella määrällä on kuitenkin haittapuolensa, sillä kaikkea julkista tietoa ei ole mahdollista kerätä (Pastor-Galindo ym. 2020). Avointen lähteiden tiedustelussa suurin haaste onkin erottaa tiedustelun kannalta relevantti tieto informaatiotulvasta (Best 2011). Ongelman ratkaisemiseksi tiedonhankintaa on enenevissä määrin ryhdytty automatisoimaan laaja-alaisen kokonaiskuvan hahmottamiseksi.

Automatisoidut tiedonhankintaohjelmat etsivät tiedustelijan antaman syötteen mukaisesti informaatiota internetistä esimerkiksi verkkosivujen metatietojen tai sisällön perusteella (Pastor-Galindo ym. 2020). Automaattinen tiedonhankinta mahdollistaa nopeasti kehittyvien tapahtumien seurannan, arvioinnin sekä reagoinnin tapahtumiin lähes reaaliajassa. Erilaisiin tiedonhankintamenetelmiin pohjautuvat automatisoidut työkalut vaihtelevat kuitenkin toimintaperiaatteiltaan, käytettävyydeltään sekä tehokkuudeltaan suuresti. Eri tiedonhankintamenetelmät soveltuvat vaihtelevasti erilaisiin tiedontarpeisiin, jolloin oikean tiedonhankintatyökalun ja -menetelmän valinta vaikuttaa merkittävästi kerätyn tiedon laatuun.

Automaattisten tiedonhankintatyökalujen soveltuvuutta on aiemmin tutkittu yksittäisten työkalujen, keräysmenetelmien ja lähteiden näkökulmasta. Tiedustelussa käytettävien työkalujen kehittämiselle ei ole olemassa yleisiä standardeja, joiden pohjalta työkaluja voisi suunnitella. Erityisesti automaattisesti kerättyjen lähteiden järjestäminen myöhempää tarkastelua varten on tunnistettu merkittäväksi haasteeksi (Best 2011). Lisäksi verkkosivustojen häviäminen sekä palveluiden rajapintojen jatkuvat muutokset vaikeuttavat luotettavan tiedon keräämistä automaattisesti (Eldridge, Hobbs ja Moran 2017). Avointen lähteiden tiedustelusta kiinnostuneet harrastajat ja ammattilaiset kehittävät alati uusia tiedonhankintatyökaluja ja -menetelmiä vastaamaan monipuolistuviin tiedontarpeisiin ja muuttuviin verkkopalveluiden rajapintoihin. Vapaasti käytettäviä työkaluja jaetaan myös aktiivisesti alaan keskittyneillä foorumeilla ja verkkosivustoilla (Bellingcat 2022; Nordine 2022).

Tässä tutkielmassa arvioidaan automaattisten tiedonhankintamenetelmien soveltuvuutta erityyppisten tiedontarpeiden tyydyttämisessä. Tiedonhankintamenetelmiä arvioidaan niiden käytettävyyden ja tehokkuuden perusteella. Tutkielman tavoitteena on tunnistaa eri tiedontarpeisiin parhaiten soveltuvat tiedonhankintamenetelmät. Lisäksi tavoitteena on tunnistaa automaattiseen tiedonhankintaan liittyviä haasteita, jotka tulisi huomioida työkalujen suunnittelussa. Tutkielma toteutetaan kirjallisuuskatsauksena, jossa keskitytään tiedustelututkimuksen ja tietotekniikan alojen tutkimuskirjallisuuteen sekä avointen lähteiden tiedusteluun erikoistuneisiin verkkojulkaisuihin.

Tutkielman näkökulma painottuu tiedusteluorganisaation suorituskyvyn kehittämiseen huomioiden tiedonhankinnan eettisen vastuun sekä operaatioturvallisuuden. Tiedustelun suorituskyvyn näkökulmasta automatisoitu tiedonhankinta vapauttaa tiedusteluorganisaation resursseja tiedonkeräyksestä tiedon käsittelyyn ja analyysiin. Tehostunut käsittely- ja analyysikapasiteetti edistää laadukkaamman tiedustelutiedon tuottamista.

Tämä tutkielma koostuu johdannosta, kolmesta sisältöluvusta ja yhteenvedosta. Ensimmäisessä sisältöluvussa esitellään tiedustelun kannalta keskeisiä käsitteitä sekä pohditaan automaation ja tiedustelun yhteensovittamisen haasteita. Toisessa sisältöluvussa syvennytään automaattisessa tiedonhankinnassa käytettäviin menetelmiin. Kolmannessa sisältöluvussa arvioidaan tiedonhankintamenetelmien soveltuvuutta eri tiedontarpeiden tyydyttämisessä.

2 Tiedustelu ja tiedonhankinta

Tässä luvussa tarkastellaan tiedon, tiedontarpeen, tiedustelun sekä avointen lähteiden tiedustelun käsitteitä ja niiden suhteita toisiinsa. Käsitteiden määrittelyn jälkeen pohditaan automaation vaikutuksia tiedonhankinnan toteuttamiseen. Ensimmäisessä alaluvussa määritellään tiedon ja tiedontarpeen käsitteet. Toisessa alaluvussa syvennytään tiedustelun olemukseen. Kolmannessa alaluvussa tarkastellaan avointen lähteiden tiedustelun ominaispiirteitä ja asemaa tiedustelulajina. Neljännessä alaluvussa vertaillaan manuaalisen ja automaattisen tiedonhankinnan eroja. Viidennessä alaluvussa pohditaan tiedustelun ja automaation yhteensovittamisen haasteita.

2.1 Tieto ja tiedontarve

Tieteellisessä tutkimuksessa tiedustelu koetaan yleensä haastavaksi aiheeksi käsitellä, sillä termille ei ole olemassa yleisesti hyväksyttyä, yksiselitteistä määritelmää. Tiedustelun määritelmää on tutkittu lähinnä tiedustelupalveluiden ja muiden julkishallinnon instituutioiden näkökulmasta (Warner 2002). Kaikkea tiedustelua yhdistää kuitenkin suhde tietoon, tiedonhankintaan ja päätöksentekoon.

Tieto käsitetään klassisessa filosofiassa 'hyvin perustelluksi, todeksi uskomukseksi'. Tiedon monitulkintaisen käsitteen vuoksi tietoa on pyritty luokittelemaan alakäsitteiksi sen välittämän arvon ja merkityksen perusteella. Rowleyn (2007) mukaan tieto voidaan jakaa hierarkkisesti dataan, informaatioon, tietämykseen ja ymmärrykseen. Alimman tason tietoa kutsutaan dataksi. Data on itsessään merkityksetöntä, kuvailevaa tietoa, jolta puuttuu konteksti ja tulkinta (Rowley 2007). Dataa ovat esimerkiksi yksittäiset symbolit, kuten numerot ja aakkoset. Informaatio on kontekstiin asetettua, järjestettyä dataa (Rowley 2007). Konteksti ja rakenne antavat datalle merkityksen, jolloin siitä tulee informaatiota. Informaatio voi olla yksittäisiä sanoja, tekstikatkelmia, kuvia tai lukuja tietokannassa. Tietämykseksi (engl. *knowledge*) kutsutaan käsitystä, joka syntyy tulkitsemalla informaatiota tietystä perspektiivistä (Rowley 2007). Tietämys on toisin sanoen asian tilaa tai siihen liittyvää informaatiota kuvaava tulkinta (Haasio ja Savolainen 2004, 17). Tulkinnan perspektiiviin vaikuttavat aiemmat käsitykset,

asenteet ja arvot. Ymmärryksellä (engl. *wisdom*) tarkoitetaan kykyä asettaa tietämystä ja informaation osia laajempaan kontekstiin sekä soveltaa kertynyttä tietämystä päätöksenteossa (Rowley 2007).

Tiedustelutoimintaan liittyvät oleellisesti myös tiedontarpeen ja tiedonhankinnan käsitteet. Tiedontarve määritellään informaatiotutkimuksessa kokemuksena toimintaympäristöön ja tilanteeseen liittyvästä epävarmuudesta (Järvelin ja Sormunen 2010, 165). Tiedontarpeen tunnistamiselle virittäviä tekijöitä ovat muun muassa käsillä oleva ongelma (Savolainen 2010, 90) sekä yksilön halu ymmärtää omaa toimintaympäristöään. Tiedontarve on siis aika- ja tilannesidonnainen ilmiö. Tarvitulla tiedolla täytyy olla pelkän tiedonhalun lisäksi käyttötarkoitus, jotta tarvittu tieto voitaisiin luokitella tiedontarpeeksi (Derr 1983). Tiedonhankinta puolestaan on määritelty tiedontarpeesta nousevaksi toiminnaksi, jonka tavoitteena on tyydyttää tiedontarve (Savolainen 2010, 91). Tiedonhankintaan kuuluu tarvittavan tiedon sisältävien lähteiden kartoittaminen, niiden luo hakeutuminen sekä niiden sisällön tulkitseminen tiedontarpeen näkökulmasta (Savolainen 2010, 91–92). Tiedonhankinnan toteutusta ohjaavat tiedontarpeen lisäksi aiemmat kokemukset tiedonhankinnasta, joiden perusteella tiedustelutehtävään valitaan soveltuvimmat tiedonhankintamenetelmät tiedontarpeen tyydyttämiseksi. Soveltuvuuteen vaikuttavat oletukset tiedonhankintamenetelmällä saatavan tiedon oikeellisuudesta, tarkkuudesta, luotettavuudesta ja määrästä.

2.2 Tiedustelu

Tiedustelutoiminnassa jalostetaan ja hyödynnetään eritasoista tietoa ymmärryksen luomiseksi. Tiedustelu voidaan käsittää ongelmanratkaisuprosessiksi, jossa kerätään, analysoidaan ja tulkitaan tietoa sekä arvioidaan toimintaympäristön olosuhteiden kehittymistä tiedon perusteella (McDowell 2009, 5). Tiedustelu on luonteeltaan soveltavaa, monialaista tutkimusta, jossa hyödynnetään tiedustelun kohteesta riippuen erilaisia tutkimusmenetelmiä ja teorioita esimerkiksi psykologiasta, yhteiskunta-, tilasto- sekä kielitieteistä.

Laajemmasta tiedonhankinnan käsitteestä poiketen tiedustelun tavoitteena on hankkia systemaattisesti tietoa päätöksenteon tueksi (Porvali 2018, 13). Tiedustelua hyödyntäviä organisaatioita ovat esimerkiksi valtionjohdot (Best 2008), viranomaiset (Bayerl ym. 2023) sekä

yritykset (Yang ja Lee 2012). Tiedustelutoimintaa voivat harjoittaa niin yksityishenkilöt kuin suuret, tiedusteluun erikoistuneet organisaatiot. Toinen tiedustelulle ominainen piirre on toiminnan salaperäinen luonne. Tiedustelun olemassaolon oikeutuksena voidaan pitää kykyä hankkia tietoa kohteista, jotka pyrkivät aktiivisesti estämään niihin kohdistuvaa tiedonhankintaa pitämällä salassa haluttuja tietoja ja peittelemällä toiminnan yksityiskohtia (Lowenthal 2020, 5). Salassa pidettyjen tietojen hankkimiseksi tiedustelu ei ilmoita sen omaamista kyvyistä, toimintamalleista, kiinnostavista kohteista eikä kerätyistä tiedoista julkisuuteen. Tällä tavoin tiedustelun kohde ei kykene tehokkaasti salaamaan arkaluontoisia tietojaan tai toimintaansa siihen kohdistuvalta tiedustelulta. Jotta tiedustelun kohde ja menetelmät eivät paljastuisi, pyritään tiedonhankinnasta jääviä jälkiä peittämään mahdollisimman hyvin jo tietoa paikannettaessa ja kerättäessä. Lisäksi tiedustelun asiakkaalle jaettava tiedustelutieto sanitoidaan eli puhdistetaan lähdeviitteistä ja tiedusteluprosessin toteutukseen liittyvistä yksityiskohdista, jotta tiedustelu ei paljastaisi tiedonhankinnassa käytettyjä lähteitä ja menetelmiä.

Tiedustelutoiminnan lähtökohtana on vastata tiedustelua toteuttavalle taholle välitettyyn tietopyyntöön, joka laaditaan havaittujen tiedontarpeiden perusteella. Tietopyynnön laatijaa kutsutaan yleensä tiedustelun asiakkaaksi tai päätöksentekijäksi (Lowenthal 2020, 263) ja tietopyynnön sisältöä kutsutaan yleisesti tiedusteluvaatimuksiksi (engl. *intelligence requirement*, JCS 2021). Tiedustelu tukee asiakkaan päätöksentekoa tuottamalla relevanttia ja ennakoivaa tiedustelutietoa toimintaympäristön muutoksiin liittyvistä riskeistä, uhista ja mahdollisuuksista (Lowenthal 2020, 2). Erotuksena laajemmasta tiedon käsitteestä tiedustelutieto on erityistä tiedontarvetta varten kerättyä, prosessoitua ja analysoitua informaatiota, jolla pyritään luomaan ymmärrys toimintaympäristön vallitsevasta tilasta (Dupont 2003). Kaikki tiedustelun keräämä tieto ei siis ole automaattisesti tiedustelutietoa, mutta kaikki tiedustelutieto on tietoa (Lowenthal 2020, 1). Tiedustelun tavoitteena on vähentää asiakkaan toimintaympäristöön liittyviä epävarmuustekijöitä, jotta tämä pystyisi ennakoimaan toimintaympäristössä tapahtuvia muutoksia. Tällöin asiakkaan on helpompi harkita vaihtoehtoisten ratkaisujen vaikutuksia toimintaympäristöön sekä tehdä saavutetun ymmärryksen perusteella paras päätös omien strategisten päämääriensä saavuttamiseksi.

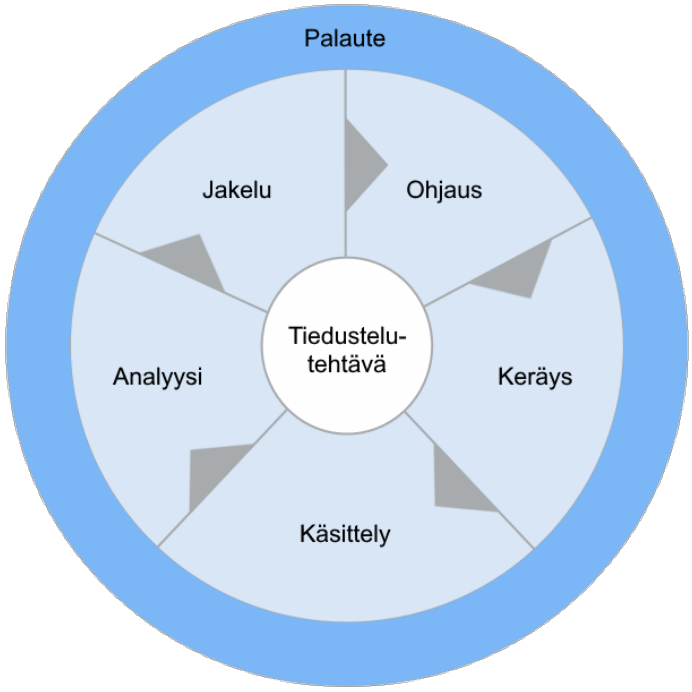
Tiedustelun viitekehyksessä tiedontarpeet liittyvät useimmiten asiakkaan toimintaympäris-

tössä vaikuttaviin toimijoihin, kuten kilpailijoihin tai vihollisiin (Warner 2002). Tiedustelun kohteita voivat olla esimerkiksi yksityishenkilöt ja useista henkilöistä koostuvat yhteisöt sekä ilmiöt, kuten epidemiat ja luonnonmullistukset. Tiedustelu pyrkii tunnistamaan toimintaympäristössä vaikuttavia toimijoita sekä arvioimaan niiden kykyjä, aikomuksia ja suhteita muihin toimijoihin (Yang ja Lee 2012). Toimintaympäristön muutokset ovat aina sidottuja aikaan, mikä muuttaa jatkuvasti päätöksentekijän tiedontarpeita.

Asiakkaiden tiedontarpeet ja intressit vaihtelevat toimialoittain. Siksi tiedusteluorganisaatiot erikoistuvat tietyn toimialan tutkimukseen, jotta tiedustelutuotteissa yhdistyisi korkea analyysiosaaminen sekä toimialalle tyypillisten tietolähteiden tuntemus. Esimerkiksi valtiojohtoa ja viranomaisia kiinnostavat yhteiskunnan vakauteen vaikuttavat ilmiöt ja toimijat (Best 2008). Liike-elämässä puolestaan yrityksiä kiinnostavat markkinatilanne, kilpailevat organisaatiot sekä kuluttajien intressit (Yang ja Lee 2012).

Tiedustelun moniulotteista olemusta voidaan siis tulkita päätöksentekoa tukevana palveluna, joka pyrkii luomaan kattavan tietopohjan päätöksenteon perustaksi ja ennakoimaan päätösten vaikutuksia. Päätöksentekijälle toimitettava tiedustelutuote syntyy tiedusteluprosessin seurauksena. Tiedusteluprosessi muodostaa tiedusteluorganisaation toiminnalle rungon, joka tukee organisaatiota laadukkaan ja käyttökelpoisen tiedustelutiedon tuottamisessa (Lowenthal 2020, 67). Prosessia havainnollistetaan usein neli-, viisi- tai kuusivaiheisen tiedustelusyklin avulla (Best 2008; Marzell 2016, 38). Syklin vaiheita suoritetaan pääosin peräkkäin ja osittain päällekkäin (Lowenthal 2020, 78–79). Tässä tutkielmassa tiedustelusyklin vaiheiksi määritellään ohjaus, keräys, käsittely, analyysi ja jakelu (kuvio 1). Lisäksi tiedustelun ja päätöksentekijän välisen vuorovaikutuksen kautta syntyvä palaute käsitetään jatkuvana toimintana tiedusteluprosessin jokaisessa vaiheessa (kuvio 1).

Tiedonhankinta käsitetään kuuluvan tiedusteluprosessin keräysvaiheeseen, joka alkaa tiedonkeräyssuunnitelman (engl. *intelligence collection plan, ICP*) laatimisella (McDowell 2009, 184–185). Tiedonkeräyssuunnitelmassa tarkennetaan tiedustelutehtävän pääpainoalueet ja käytettävät tiedonhankintamenetelmät. Tiedonhankintamenetelmät valitaan kohteen ominaisuuksien mukaan siten, että kohteesta saadaan hankittua mahdollisimman tarkkaa ja relevanttia informaatiota minimoiden samalla epäoleellisen informaation kerääminen (McDowell 2009, 184). Keräysvaiheessa löydetyn informaation relevanssia arvioidaan pintapuolisesti,



Kuvio 1. Havainnekuva tiedustelusyklin rakenteesta ja tiedusteluprosessin kulusta. Mukailtu Yhdysvaltain puolustushaarakomentajien neuvoston (2019) julkaisusta.

minkä jälkeen oleelliseksi katsottu informaatio tallennetaan myöhempää käsittelyä varten. Vaikka tiedonhankinta toteutetaan pääosin tiedonkeräyssuunnitelman mukaisesti, kohdennetaan tiedonhankintaa kerätystä informaatiosta tehtyjen arvioiden perusteella (McDowell 2009, 183). Näin maksimoidaan mahdollisuudet löytää relevanttia tietoa tarkentamalla tiedonhaun parametreja tai käytettyjä menetelmiä. Automatisoimalla keräysvaiheen tehtäviä tiedustelu voi keskittää henkilöstöä tiedusteluprosessin käsittely- ja analyysivaiheisiin, joissa korostuu asiantuntijaosaamisen tärkeys.

2.3 Avointen lähteiden tiedustelu

Avointen lähteiden tiedustelu (engl. *open-source intelligence, OSINT*) on tiedustelulaji, joka keskittyy keräämään tietoa julkisesti saatavilla olevista lähteistä (Best 2008). Tässä tutkielmassa avointen lähteiden tiedustelusta käytetään jatkossa lyhennettä OSINT. OSINT-tietoja voidaan kerätä kaikista lähteistä, joihin on mahdollista päästä käsiksi laillisin keinoin. Lail-

lisiin keinoihin lukeutuvat pääsyn pyytäminen tai ostaminen tietoon sekä tiedustelijan omakohtainen havainnointi vapaasti saatavilla olevasta informaatiosta (Jardines 2016, 5). Lailliset keinot rajaavat OSINTin ulkopuolelle luvattoman tunkeutumisen, varkauden, hakkeroinnin ja sosiaalisen manipuloinnin eli ihmiseen kohdistuvan vihamielisen vaikuttamisen kautta saatavan tiedon (Jardines 2016, 5).

Avoimista lähteistä kerättyä tietoa luokitellaan yleensä tiedon rakenteisuuden, jalostusasteen, formaatin tai jakelutavan perusteella. Tiedon rakenteisuudella tarkoitetaan tiedon osien keskinäisen järjestyksen tarkkuutta ja loogista jäsennystapaa (Gibson 2016, 73). Rakenteista tietoa ovat esimerkiksi kartastot ja tietokantojen järjestetyt tietovarannot. Rakenteetonta tietoa puolestaan ovat esimerkiksi tekstitiedostot ja kuvat. NATO (2002) luokittelee kerätyt tiedot niiden jalostusasteen mukaan dataksi, informaatioksi, tiedustelutiedoksi ja vahvistetuksi tiedustelutiedoksi. Tiedon formaatti ja jakelutapa määrittävät vahvasti menetelmät, joilla tietoa on mahdollista etsiä ja kerätä. OSINT-tieto voi olla formaatiltaan esimerkiksi luonnollista tekstiä, kuvia, ääntä, videokuvaa tai tietoaineistoja (Pastor-Galindo ym. 2020). OSINT-tietoa voidaan luokitella jakelutavan perusteella massamedian lähetyksiin, painettuihin julkaisuihin sekä digitaaliseen tietoon (Jardines 2016, 22–28).

OSINT on noussut kirjallisten ja digitaalisten aineistojen voimakkaan lisääntymisen myötä edulliseksi ja helposti lähestyttäväksi tiedustelulajiksi. Internetin käyttö massamedian julkaisukanavana sekä painettujen aineistojen digitointi ovat helpottaneet yksityishenkilöiden pääsyä avoimen tiedon äärelle (Benes 2013). OSINT soveltuu erinomaisesti tilannekuvan muodostamiseen toimintaympäristöstä sekä perustietojen keräämiseen heikosti tunnetuista kohteista. Esimerkiksi viranomaiset ja muut valtiolliset toimijat hyödyntävät OSINT-menetelmiä massamedioiden reaaliaikaisessa seurannassa kriisien ennakoimiseksi (Best 2008). OSINT soveltuu myös tiedustelun asiakkaan tekemien päätösten vaikuttavuuden arviointiin lyhyellä aikavälillä (Eldridge, Hobbs ja Moran 2017). Lähteiden julkinen luonne edistää päätöksenteon läpinäkyvyyttä tilanteissa, joissa OSINT-tietoja käytetään perusteena päätöksille (Pastor-Galindo ym. 2020). Vaikka tiedustelutiedot ovat lähtökohtaisesti aina salaisia, voidaan OSINT-tietoja jakaa tiedustelun asiakkaalle matalammalla kynnyksellä ja vähäisemmällä sanitoinnilla kuin muilla tiedustelumenetelmillä kerättyjä tietoja. Läpinäkyvyys lisää sidosryhmien luottamusta päätöksentekijää ja tiedustelua kohtaan sekä mahdollistaa lähtei-

den omakohtaisen etsimisen ja arvioinnin päätöksen perusteiden oikeellisuuden varmistamiseksi.

Avoimista lähteistä kerättävien tietojen arvioidaan muodostavan noin 80 % kaikesta tiedustelutiedosta, jota valtion viranomaiset hyödyntävät päätöksenteossaan (Dupont 2003; Hulnick 2002). Tiedustelutoiminnan salaisen luonteen takia OSINT-tietojen tarkkaa osuutta kaikesta tiedustelutiedosta ei ole mahdollista vahvistaa. Muihin tiedustelulajeihin verrattuna OSINT on edullinen ja nopea tapa hankkia tietoa useista eri lähteistä (Dupont 2003; Ünver 2018). Kalliisiin sensorijärjestelmiin sijoittamisen tai vuosia kestävän henkilöstökoulutuksen sijaan OSINTiin vaaditaan vain päätelaite ja verkkoyhteys. OSINTin edullisuus antaa tiedusteluorganisaatiolle mahdollisuuden keskittää taloudellisia resurssejaan esimerkiksi laadukkaampaan henkilöstökoulutukseen tai muihin tiedustelulajeihin liittyviin kalliisiin investointeihin.

Muihin tiedustelulajeihin verrattavana heikkoutena OSINTin avulla kerätty tieto on aina toisen käden tietoa (Jardines 2016, 6). Tiedon toiskätisyydellä tarkoitetaan, että OSINT-tiedustelija ei itse havaitse tutkittavaa ilmiötä ja siihen liittyvää informaatiota, vaan kerää jonkun toisen havaitsemaa, koostamaa ja julkaisemaa tietoa. Toisaalta julkisesti levitetty, jalostettu tieto voi helpottaa tiedustelijan työtä, sillä valmiiksi koostettu tieto vaatii yleensä vähemmän käsittelyä kuin useista eri lähteistä kerätty raaka tieto (Lowenthal 2020, 136). Toisen käden tieto ei itsessään ole vähempiarvoisempaa kuin ensi käden tieto, mutta sen alkuperäisen kerääjän ja julkaisijan motiiveja on pohdittava tiedon luotettavuuden arvioinnin kannalta. Toisen käden tieto voi olla vahingossa tai tahallisesti levitettyä virheellistä tietoa eli mis- tai disinformaatiota. Tiedon ja lähteiden luotettavuuden arviointi onkin yksi ammattimaisen OSINT-toiminnan kulmakivistä (Hulnick 2002).

2.4 Tiedonhankintamenetelmät avointen lähteiden tiedustelussa

Tiedonkeräyssuunnitelman laatimisen ja pohjatietojen kartoittamisen jälkeen tiedonhankintaa ryhdytään toteuttamaan tiedustelun kohteesta tunnettujen lähtötietojen perusteella. Lähtötietoja voivat olla esimerkiksi nimet, nimimerkit, sähköpostiosoitteet, sijaintitiedot tai IP-osoitteet (Pastor-Galindo ym. 2020). Tiedonhaun perusteella löytyvää relevanttia tietoa kerätään ja analysoidaan karkeasti, jotta tiedonhaussa käytettäviä hakutermejä voitaisiin tarken-

taa tiedonhaun kohdentamiseksi (McDowell 2009, 183).

OSINT voidaan jakaa tiedonhankinnan toteutustavan perusteella manuaaliseen, passiiviseen ja automaattiseen tiedonhankintaan (Bazzell 2019, 516–518). Manuaalinen tiedonhankinta on tiedustelijan suorittamaa tiedonhankintaa, jossa tämä käyttää erilaisia tiedonhankinnan työkaluja päästäkseen käsiksi lähteeseen ja havainnoidakseen lähteen sisältämää informaatiota. Tiedustelija käyttää tiedonhankinnassa harkittuja hakusanoja ja -tekniikoita löytääkseen tiedontarpeen kannalta relevanttia informaatiota (Best 2008). Kerätessään tietoa tiedustelija arvioi tiedon oikeellisuutta sekä relevanssia tiedontarpeen näkökulmasta. Tiedustelija tallentaa oleellisiksi arvioidut tiedot prosessointia ja analyysia varten (Bazzell 2019, 517). Manuaalista tiedonhankintaa voidaan tukea passiivisilla tiedonhankintatyökaluilla. Passiivinen tiedonhankinta on tietokoneohjelman suorittamaa tiedonkeräystä manuaalisen tiedonhankintaprosessin kulusta (Bazzell 2019, 517). Esimerkiksi tiedustelijan vierailemien verkkosivustojen metatietojen sekä vierailujärjestyksen tallentaminen lokiin ovat passiivisen tiedonhankinnan tehtäviä.

Manuaalisen ja passiivisen tiedonhankinnan ohella OSINT-tietoja voidaan paikantaa ja kerätä automatisoidusti digitaalisilta alustoilta. Automaattinen tiedonhankinta on tietokoneohjelmalle ennalta määriteltyjen komentosarjojen ja tiedustelijan syöttämien parametrien perusteella suoritettavaa tiedonhankintaa (Bazzell 2019, 518). Tiedonhankintaohjelmat hyödyntävät internetin rakennetta, lähteeseen ja sen sisältämään informaatioon liittyviä metatietoja sekä erilaisia hakumenetelmiä kiinnostavan tiedon paikantamiseksi ja poimimiseksi (Gibson 2016, 84). Automaattisesta tiedonhankinnasta poiketen passiivista tiedonhankintaa suoritetaan vain manuaalisen tiedonhankintaprosessin aikana. Automaattista tiedonhankintaa voidaan puolestaan suorittaa täysin erillään manuaalisesta tiedonhankinnasta.

Automaation soveltaminen OSINTissa siirtää pitkästyttävät ja vaativuustasoltaan yksinkertaiset tehtävät tiedustelijoilta ohjelmistoille, jolloin tiedustelijoiden asiantuntijaosaamista voidaan hyödyntää tiedusteluprosessin muissa vaiheissa (Bayerl ym. 2023). Automaattisen tiedonhankinnan suurimpia etuja ovat tiedonkeräyksen nopeus ja jatkuvuus sekä potentiaalisesti suuri kerätyn tiedon määrä. Tiedusteluprosessin aikana kerätyn tiedon määrän on havaittu vaikuttavan tiedusteluanalyysin tarkkuuteen (Pastor-Galindo ym. 2020). Suurempi relevantin lähdetiedon määrä auttaa tunnistamaan tiedustelun kohteeseen liittyviä toimijoita sekä

kohteen ja toimijoiden välisiä suhteita (Eldridge, Hobbs ja Moran 2017). Internetin valtavan tietomäärän takia kaikkea tietoa ei ole mahdollista kerätä tarkan kokonaiskuvan muodostamiseksi (Pastor-Galindo ym. 2020). Automatisoidut tiedonhankintamenetelmät mahdollistavat kuitenkin laajemman tietomassan keräämisen manuaaliseen tiedonhankintaan verrattuna. Tämä puolestaan mahdollistaa pinnalla olevien ilmiöiden ja laajempien trendien havaitsemisen informaatiovirrasta (Hoppa ym. 2020).

Automaattisten tiedonhankintaohjelmien merkittävänä etuna on inhimillisten tekijöiden ja virheiden minimointi tiedustelussa. Tiedonhankintaohjelmat eivät ole ihmisten tavoin alttiita kognitiivisille ajattelun vinoumille, jotka saattavat alitajuisesti ohjata tiedonhankintaa (Ponder-Sutton 2016, 11–12). Manuaalisen tiedonhankinnan tehokkuuteen vaikuttavat tiedustelijan kokemus, tekniset taidot, vireystila, motivaatio ja monet muut tekijät. Manuaaliseen tiedonhankintaan verrattuna automaattiset tiedonhankintaohjelmat kykenevät käsittelemään havaittuja tietoja tasaisella intensiteetillä, johon vaikuttavat ainoastaan ohjelmaa suorittavan tietokoneen laskentateho sekä verkon tiedonsiirtonopeus. Vaikka automaattiset tiedonhankintaohjelmat eivät kykene toimimaan täysin autonomisesti ilman käyttäjäsyötettä, voivat työkalut jatkaa tiedonhankintaa annetun syötteen rajoissa käytännössä loputtomiin (Eldridge, Hobbs ja Moran 2017). Tästä on hyötyä jatkuvan tilannekuvan muodostamisessa, jota automaattinen tiedonhankinta tukee merkittävästi. Tiedonhankintaohjelmat kykenevät suorittamaan tiedonkeräystä mihin vuorokaudenaikaan tahansa ja seuraamaan useiden eri lähteiden julkaisuja herkemmin kuin ihminen.

2.5 Tiedustelun ja automaation haasteet

Tiedonhankintaan liittyy lukuisia haasteita, jotka vaativat jatkuvaa toiminnan arvioimista tiedusteluprosessin sujuvuuden sekä tiedustelutiedon laadun säilyttämiseksi. Automaatio ei ratkaise tiedustelun luontaisia haasteita, vaikka se tehostaa tiedon hankinta- ja käsittelykykyä. Automaation käyttö tiedustelussa voi pahimmassa tapauksessa vaikeuttaa tiedustelutyötä, jos automatisoituja ratkaisuja suunnitellaan harkitsemattomasti.

Tiedonhankintatyökalujen kehittäminen ja testaaminen vie aikaa, jolloin yksittäisiä tiedustelutehtäviä varten tarvittavien työkalujen kehittäminen ei yleensä ole järkevää pitkällä täh-

täimellä. Tämän vuoksi tiedonhankintatyökalujen tulisi olla tarpeeksi joustavia käytettäväksi erilaisissa tiedustelutehtävissä, mutta tarpeeksi erikoistuneita, jotta niillä voitaisiin kerätä tarkkaa ja relevanttia tietoa kohteista.

2.5.1 Tiedon määrä ja relevanssi

Tiedonhankinnassa on ensiarvoisen tärkeää löytää relevanttia tietoa tiedontarpeen tyydyttämiseksi. Relevanssilla tarkoitetaan informaation arvioitua käyttökelpoisuutta tietyssä käyttötilanteessa ottaen huomioon tiedon käyttäjän tavoitteet, arvot ja odotukset (Järvelin ja Sormunen 2010, 164–165). Ennen tiedonhankinnan aloittamista tiedustelun tulisi olla perillä tiedustelutehtävälle asetetuista tiedusteluvaatimuksista, joihin ryhdytään etsimään tarkkoja vastauksia (Lowenthal 2020, 67). Epätarkat tiedusteluvaatimukset johtavat usein epärelevantin tiedon keräämiseen, mikä ei palvele asiakasta eikä päätöksentekoprosessia (Martelius 2020).

Vaikka lähdetiedon saatavuus on yksi OSINTin merkittävimmistä vahvuuksista, on tiedon valtavan määrän hallitseminen keskeinen haaste tiedustelulle (Best 2011; Jardines 2016, 26). Suurista tietomassoista kerätyn informaation tarkastelu mahdollistaa trendien erottamisen informaatiovirrasta (Hoppa ym. 2020). Samalla trendeihin keskittyminen saattaa estää tiedustelijaa havaitsemasta tärkeitä yksityiskohtia tai poikkeamia kerätyn informaation seasta (Eldridge, Hobbs ja Moran 2017). Kriittisten yksityiskohtien havaitsematta jättäminen voi koitua kohtalokkaaksi etenkin, jos tiedustelun päämääränä on tarkkailla potentiaalisia turvallisuusuhkia.

Saatavilla olevan tiedon valtava määrä vaikeuttaa oleellisesti relevantin tiedon löytämistä tietovirrasta (Eldridge, Hobbs ja Moran 2017). Harjaantunut tiedustelija kykenee arvioimaan informaation relevanssia manuaalisen tiedonhankinnan aikana, jolloin epäoleellisia tietoja voidaan karsia ennen käsittelyvaihetta. Tiedonhankintaohjelmat eivät kuitenkaan voi päätellä tarkasti, vastaako löydetty tieto tarkasti tiedontarpeeseen. Automatisoitujen tiedonhankintatyökalujen suunnittelussa ja käytössä tulisi ottaa huomioon syötteen parametrien monipuolisuus ja tarkkuus, jotta relevanttia informaatiota voitaisiin löytää nopeasti. Relevantin informaation erottelukyky heijastuu hankitun tiedon määrän lisäksi tiedusteluanalyysin laa-

tuun (Pastor-Galindo ym. 2020). Keräyskapasiteetin kasvaessa myös tiedontarpeen kannalta oleellisen tiedon löytäminen vaikeutuu, koska kerättyä informaatiota on enemmän käsiteltäväksi. Tämän takia tiedonhankinnan tuloksia tulisi arvioida jo keräysvaiheessa, jotta tiedonhankintaohjelmien syötteitä voitaisiin päivittää ja tarkentaa parempien lähteiden ja tiedon löytämiseksi.

Tiedustelun tehtävä- ja kontekstisidonnaisuus vaativat tiedonhankintamenetelmien soveltamista kuhunkin tehtävään sopivalla tavalla, jolloin kerätyt lähteet voivat myös vaihdella suuresti. Kaikki informaatio ei lähtökohtaisesti ole yhtä arvokasta, minkä vuoksi tiedonhankinnassa tulisi keskittyä tiedontarpeiden tyydyttämisen ohella laadukkaiden lähteiden etsimiseen (McDowell 2009, 182; Lowenthal 2020, 73). Monesti laadukkaat lähteet sisältävät tiedusteluvaatimusten lisäksi paljon informaatiota, jota voidaan hyödyntää myöhemmissä tiedustelutehtävissä. Siksi laadukkaiksi katsotut lähteet tulisi tallentaa ja indeksoida tiedusteluorganisaation tietovarantoihin, jotta tiedonkeräystä ei tarvitsisi suorittaa tarpeettomasti uudelleen samankaltaisen tiedustelutehtävän kohdalla.

Lähdetiedon määrää kasvattaa myös tiedon toisteisuus. Toisteisella tiedolla tarkoitetaan useassa eri lähteessä julkaistua tietoa, joka pohjautuu samaan ilmiöön tai havaintoon (Lowenthal 2020, 137). Esimerkiksi uutisartikkelit saattavat sisältää täsmälleen saman informaation kuin muut aiheesta julkaistut artikkelit. Toisteisen tiedon suuri määrä voi vinouttaa käsitystä tutkitun ilmiön merkittävyydestä. Automaattisissa tiedonhankintatyökaluissa hyödynnetään jäsenninalgoritmeja, jotka tunnistavat toisteisen tiedon ja käsittelevät eri lähteistä saadun saman tietosisällön yksittäisenä tiedusteluhavaintona (Best 2011). Toisteista tietoa julkaisevien lähteiden seuraaminen voi lisäksi paljastaa lähteiden takana olevien organisaatioiden motiiveja sekä niiden käyttämiä tietolähteitä (Jardines 2016, 31).

Julkisen tiedon suurta määrää ei tulisi käyttää oletuksena sille, että tiedontarpeeseen on olemassa yksiselitteinen, muut vaihtoehdot poissulkeva vastaus (Weir 2016). Kerätyistä tiedoista koottu kokonaiskuva on lähes aina puutteellinen, sillä kaikkea tutkittavaan ilmiöön liittyvää tietoa ei ole julkisesti saatavilla tai edes mahdollista hankkia. Jos manuaalisella tiedonhankinnalla ei ole mahdollista löytää vastausta tiedontarpeeseen, on se myös mahdotonta automatisoiduin keinoin (Weir 2016). Toisaalta automaattisilla menetelmillä kerättävissä olevaa tietoa voi rajoittaa vaatimus maksaa tietolähteen käytöstä tai verkkopalveluun tunnis-

tautumisesta. Automaattisella tiedonhankinnalla voisi tosin selvittää nopeasti, onko tiedontarpeisiin saatavissa tyydyttäviä vastauksia avoimista lähteistä. Vastausten alustavan saatavuuden selvittäminen voi säästää tiedustelun resursseja kohdentamalla ne muihin tiedustelukohteisiin.

2.5.2 Tiedon ja lähteiden luotettavuus

Toinen OSINT-menetelmillä kerättyyn tietoon liittyvä ongelma on tiedon oikeellisuuden ja lähteen luotettavuuden arviointi (Pastor-Galindo ym. 2020; Weir 2016). Verkossa kuka tahansa voi julkaista sisältöä anonyymisti, minkä vuoksi informaation ja lähteisiin tulisi aina suhtautua epäilevästi. Virheellinen tieto voi olla tahattomasti levitettyä tietoa eli misinformaatiota tai tahallisesti harhaanjohtavaa tietoa eli disinformaatiota. Mis- ja disinformaatio ovat tiedon oikeellisuuden arvioimisen lisäksi haaste lähteen luotettavuuden arvioinnille, sillä lähteen motiivien arviointi on usein vaikeaa (Weir 2016).

Tarkasteltavan tiedon oikeellisuuden arvioimiseksi tietoa hankitaan useista toisistaan riippumattomista lähteistä, jolloin tarkasteltavan tiedon uskottavuus vahvistuu (Weir 2016). Tämä ei kuitenkaan tee arvioidusta tiedosta automaattisesti paikkansapitävää, vaan ainoastaan nostaa sen todennäköisyyttä olla paikkansapitävää. Lähteiden luotettavuuden arvioimiseksi käytetyistä tietolähteistä pidetään yleensä kirjaa, jotta tietojen oikeellisuutta voitaisiin arvioida jälkikäteen. Tiedusteluorganisaatiot kykenevät pitkän aikavälin seurannalla asettamaan lähteet luotettavuuden mukaan paremmuusjärjestykseen, mikä edistää lähteestä kerättävien tietojen arviointia tulevaisuudessa (Gibson, Ramwell ja Day 2016, 107).

OSINT-tiedustelijat oppivat kokemuksen ja kehittyvän kohdeosaamisen kautta tunnistamaan epäluotettavia lähteitä sekä suoranaista disinformaatiota (Hulnick 2002). Harjaantunut tiedustelija kykenee karsimaan epäluotettavia lähteitä jo manuaalisen keräysvaiheen aikana, jolloin käsittelyvaiheen resursseja säästyy. Automaattisessa tiedonhankinnassa tulisikin kiinnittää huomiota ohjelman kykyyn arvioida lähteen semantiikkaa. Sisällön semantiikan tunnistamisen lisäksi olisi tärkeätä kyetä tunnistamaan lähteen lähestymistapaa ja asenteita tietoja kohtaan. Pakinoiden, blogien ja tiedeartikkeleiden erottelu erillisiksi tekstilajeiksi helpottaa lähteen ja tiedon luotettavuuden arviointia keräys- ja käsittelyvaiheessa.

3 Automaattiset tiedonhankintamenetelmät

Tässä luvussa käsitellään automaattisen tiedonhankinnan toteuttamista verkossa. Aluksi esitellään automaattisen tiedonhankinnan vaiheet, jonka jälkeen perehdytään tarkemmin tiedonhankinnassa käytettäviin tiedonpaikannus- ja tiedonkeräysmenetelmiin.

Verkossa tapahtuva tiedonhankintaprosessi voidaan jakaa kolmeen osaan. Ensimmäiseksi on paikannettava verkkosivusto, joka sisältää tiedontarpeen kannalta relevanttia informaatiota. Verkkosivuston paikantamisen jälkeen sen kanssa tulee kyetä vuorovaikuttamaan, jotta informaatiosisältöä voisi käsitellä. Lopuksi verkkosivuston sisältämää informaatiota on kyettävä arvioimaan tiedontarpeen perusteella annetun syötteen perusteella ja poimimaan oleellinen tieto verkkosivustolta.

Tässä tutkielmassa tiedonhankintamenetelmät ymmärretään tiedonpaikannus- ja tiedonkeräysmenetelmien yhdistelminä. Tiedon paikannuksella tarkoitetaan relevanttia informaatiota sisältävän tietolähteen etsimistä ja paikantamista internetistä sekä tietoon käsiksi pääsyä. Tiedonkeräyksellä tarkoitetaan relevantin informaation arvioimista, poimimista tietolähteestä ja tallentamista myöhempää käsittelyä varten. Paikannus- ja keräysmenetelmien välille on haastavaa tehdä selkeätä rajanvetoa, koska ne usein hyödyntävät toistensa kykyjä käsitellä lähteitä ja informaatiota.

3.1 Tiedonpaikannusmenetelmät

Tiedonpaikannusmenetelmät ovat automaattisessa tiedonhankinnassa vastuussa relevantin tiedon löytämisestä tiedonhankintaohjelmalle annetun syötteen perusteella. Tässä alaluvussa tarkastellaan verkkosyötteiden, ohjelmointirajapintojen ja hakurobottien käyttömahdollisuuksia tiedon paikantamiseksi verkosta.

3.1.1 Verkkosyötteet

Verkkosyöte (engl. *web feed*) on tiedonjakelukanava, jossa verkkosivuston ylläpitäjä voi ilmoittaa syötteen tilaajille sivuston päivityksistä ja uusista julkaisuista. Ilmoitukset jaetaan

verkkosyötteen tilaajille yleensä XML-kielisenä tiedostona, johon on koostettu viimeisimmät tiedot verkkosivuston muutoksista (Urbansky ym. 2021). Syötteen tilaaminen perustuu verkko-osoitteeseen lähetettävään päivityspyyntöön, johon verkkosyötettä ylläpitävä palvelin vastaa viimeisimmät päivitykset sisältävällä tiedostolla (Gibson 2016, 81). Palvelimille lähetettäviä päivityspyyntöjä voidaan automatisoida syötteenlukuohjelmien avulla, jotka pitävät kirjaa seuratuista verkkosyötteistä ja koostavat syötteissä ilmoitetut muutokset tiedustelijalle luettavaksi (Urbansky ym. 2021). Erityisesti uutispalvelut, blogit ja sosiaalisen median kanavat suosivat verkkosyötteitä, sillä niiden kautta lukijakunta on mahdollista tavoittaa nopeasti.

Verkkosyötteiden vahvuuksia ovat niiden helppokäyttöisyys, rakenteellisen tiedon saatavuus sekä reaaliaikaisuus. Verkkosyötteen ylläpitäjä koostaa ilmoituksen standardoituun tiedostomuotoon, mikä tekee tiedon käsittelystä ja analysoinnista tiedustelulle ketterämpää. Esimerkiksi Euroopan Unionin (2023) ylläpitämä European Media Monitor -palvelu seuraa aktiivisesti uutispalveluiden verkkosyötteitä mahdollistaen kehittyvien tilanteiden seuraamisen reaaliajassa. Ajantasainen informaatio parantaa huomattavasti tilannekuvan muodostamista toimintaympäristöstä. Informaation reaaliaikaisuus perustuu kuitenkin palvelimelle lähetettävien pyyntöjen tiheyteen, jolloin uuden informaation haku verkkosyötteistä on tiedustelun ja tiedonhankintaohjelmien vastuulla (Urbansky ym. 2021).

Verkkosyötteiden haittapuolia ovat niiden päämäärähakuinen ja verrattain vähäinen käyttö tiedonjakelukanavana. Palvelun ylläpitäjät kontrolloivat syötteessä jaettuja tietoja, jolloin syötteessä julkaistaan harvoin tahattomasti arkaluontoista informaatiota. Verkkosyötteet soveltuvat myös kohdennettuun tiedonhankintaan huonosti, sillä niissä julkaistaan tietoa lähinnä ajankohtaisista tapahtumista.

3.1.2 Ohjelmointirajapinnat

Ohjelmointirajapinta eli API (engl. *application programming interface*) on verkkosivuston tarjoama palvelu, joka mahdollistaa sivuston ylläpidon ulkopuolisille käyttäjille pääsyn verkkosivuston sisäänrakennettuihin toimintoihin (Perez ja Germon 2015, 110). Verkkopalvelun käyttäjät voivat lähettää rajapinnan kautta ennalta määriteltyjä kyselyjä suoraan palvelimel-

le, joka palauttaa vastauksen rakenteisessa muodossa käyttäjälle (Perez ja Germon 2015, 110). Ohjelmointirajapintoja käytetään yleensä eri ohjelmistojen välisessä tiedonvälityksessä, mutta niiden kautta voidaan hankkia verkkosivustolta helposti saatavissa olevaa informaatiota (Chaudhary ja Bansal 2022). Esimerkiksi monet sosiaalisen median palvelut ja uutissivustot sallivat ohjelmointirajapintansa kautta yksittäisten julkaisujen tai käyttäjätilien hakemisen. Julkaisuissa näkyvän informaation lisäksi jotkut ohjelmointirajapinnat mahdollistavat myös julkaisun metatietojen hakemisen.

Ohjelmointirajapinta on tehokas tiedonpaikannusmenetelmä kohdistetussa tiedonhankinnassa. Uutisartikkelit ja sosiaalisen median julkaisut noudattavat usein rakennetta, jossa julkaisuun liittyvät tiedot, kuten tekstisisältö, julkaisijan nimimerkki, julkaisuajankohta ja sijainti ovat upotettuina omiin kenttiinsä verkkosivustolla (Chaudhary ja Bansal 2022). Näihin tietoihin voi useimmiten päästä suoraan käsiksi ohjelmointirajapinnan kautta, mikä mahdollistaa relevanttien julkaisujen ja käyttäjätilien etsimisen syötettyjen parametrien avulla. Ohjelmointirajapintojen käyttöä voidaan automatisoida lähettämällä kyselyjä useisiin verkkopalveluihin samoilla parametreilla.

Ohjelmointirajapintojen heikkoutena ovat verkkosyötteiden tavoin palveluntarjoajan asettamat rajoitteet, joiden puitteissa rajapintoja voidaan hyödyntää tiedonhankinnassa. Kaikkeen verkkopalvelun sisältöön ei välttämättä pääse käsiksi pelkkien ohjelmointirajapintojen avulla. Lisäksi jotkut ohjelmointirajapinnat vaativat käyttäjältä tunnistautumista palvelun toiminnallisuuksien käyttämiseksi (Perez ja Germon 2015, 110). Tunnistautuminen toteutetaan yleensä API-avaimella, joka on käyttäjälle erikseen luotu uniikki merkkijono, joka liitetään palvelimelle lähetettävään kyselyyn. Palveluntarjoajat kykenevät hallitsemaan API-avaimilla palveluidensa käyttöä sekä ehkäisemään palvelinten ylikuormitusta. Käyttäjän API-avaimen perusteella palveluntarjoaja voi rajoittaa ohjelmointirajapinnan maksullisten toimintojen käyttöä (Gibson 2016, 77) ja ehkäistä palvelimeen kohdistuvaa kuormitusta rajoittamalla yksittäisellä API-avaimella tehtävien pyyntöjen määrää (Perez ja Germon 2015, 110). Palveluntarjoajalla on API-avaimien myötä mahdollisuus seurata kullakin avaimella tehtyjen kyselyjen sisältöä, mikä muodostuu ongelmaksi tiedonhankinnan salaamisessa.

3.1.3 Hakurobotit

Hakurobotti (engl. *web crawler, web spider*) on tiedonhakuohjelma, joka vierailee järjestelmällisesti eri verkkosivustoilla kartoittaen niiden rakennetta ja yhteyksiä muihin verkkosivustoihin. Esimerkiksi verkkoselainten hakukoneiden valtavat hakemistot perustuvat lukuisiin hakurobotteihin, jotka kartoittavat internettiä automaattisesti uusien verkkosivustojen liittämiseksi hakukonetuloksiin (Hassan ja Hijazi 2018, 96–97). Hakurobotteja voidaan käyttää kohdistetussa tiedonhankinnassa, jossa ne kartoittavat verkkosivustoja annetun syötteen perusteella.

Hakurobotin toiminta perustuu hyperlinkkien tunnistamiseen verkkosivuston rakenteesta (Gibson 2016, 75). Haku aloitetaan syöttämällä hakurobotille siemenlinkki (engl. *seed URL*), joka ohjaa robotin ensimmäiselle kartoitettavalle sivustolle (Chaudhary ja Bansal 2022). Siemenlinkin lisäksi määritetään haun syvyys, jolla tarkoitetaan etäisyyttä, kuinka monen verkkosivuston päähän siemenlinkistä hakurobotti vierailee (Gibson 2016, 75). Hakurobotti skannaa sivuston hyperlinkeistä, jotka robotti tallentaa listalle. Tämän jälkeen hakurobotti vierailee ja toistaa toimenpiteet rekursiivisesti jokaiselle listan hyperlinkille, kunnes se saavuttaa haun syvyyden (Chaudhary ja Bansal 2022). Perinteisen hyperlinkkejä seuraavan hakurobotin lisäksi robotti voidaan rakentaa yhdistäen verkkoselaimen hakukoneiden toiminnallisuuksia. Hakurobotin syötteenä voidaan tällöin käyttää tavallisen hakukonehaun tavoin avainsanoja, joiden perusteella hakukone etsii ja päättää verkkosivustot, joilla hakurobotti vierailee (Di Pietro ym. 2014).

Verkkosivustoilla vierailemisen lisäksi hakurobottiin on yleensä liitettynä tiedonkeräysohjelma, joka poimii ja käsittelee verkkosivuston sisältöä relevantin informaation löytämiseksi (Chaudhary ja Bansal 2022). Tiedonkeräysohjelmat poimivat esimerkiksi tekstiä, kuvia ja metatietoja, joiden avulla voidaan päätellä sisällön konteksti ja arvo tiedustelulle (Aliprandi ym. 2014). Kohdennetussa tiedonhaussa hakurobotit hyödyntävät tiedonkeräysmenetelmiä arvioidakseen, sisältääkö kartoitettava verkkosivusto tiedontarpeisiin nähden relevanttia informaatiota (Gibson 2016, 75–76).

Hakurobotit ovat eteviä perustietojen selvittämiseksi tiedustelun kohteesta. Toisin kuin verkkosyötteet ja ohjelmointirajapinnat, hakurobotit eivät ole riippuvaisia verkkosivustojen yllä-

pitäjien tarjoamista toiminnoista, vaan ne toimivat täysin oman sovelluslogiikkansa mukaan. Hakurobotteihin liitettävät tiedonkeräysohjelmat ja niiden muokkaaminen tekevät hakuroboteista joustavia työkaluja suurten verkkosivustojen ja tietokantojen kartoittamisessa. Esimerkiksi sosiaalisen median palveluissa hakurobotti voi automaattisesti kartoittaa tiettyyn käyttäjätiliin linkitetyt tilit ja analysoida tilien välisiä suhteita (Aliprandi ym. 2014).

Hakurobottien merkittävistä eduista huolimatta niiden toimintaan liittyy myös haasteita. Ensimmäisenä haasteena on hakurobottien suhteellinen hitaus tiedonhankinnassa. Verkkosivustojen kartoittamiseen kuluu paljon aikaa, koska hakurobotin tulisi vierailun lisäksi kahlata läpi verkkosivuston sisältö ja analysoida sen relevanssia kiinnostavan informaation paikantamiseksi (Ponder-Sutton 2016, 9). Potentiaalisesti relevanttien lähteiden löytyessä verkkosivuston sisällön lataaminen ja tallentaminen vie myös aikaa, jos sivusto sisältää suurikokoisia tiedostoja, kuten kuvia tai videoita.

Vaikka hakurobottien toiminta perustuu niiden kehittäjän laatimiin algoritmeihin, voivat verkkosivuston ylläpitäjät ohjata hakurobottien toimintaa. Ylläpitäjä voi määritellä *robots.txt*-nimisessä tiedostossa verkkosivustoilla vieraileville hakuroboteille kartoitettavaksi sallitut kansiot ja tiedostot (Gibson 2016, 91). Mikäli hakurobotti sivuuttaa *robots.txt*-tiedoston käsittelyn, voi verkkosivuston ylläpitäjä estää robotilta kokonaan pääsyn sivustolle (Gibson 2016, 91).

3.2 Tiedonkeräysmenetelmät

Automaattisessa tiedonhankinnassa tiedonkeräyksellä tarkoitetaan relevantin informaation eristämistä, poimimista ja tallentamista lähteestä. Tiedonkeräystä, joka kohdistuu nimenomaan internet-lähteisiin, kutsutaan myös verkkoharavoinniksi (engl. *web scraping*) (Diouf ym. 2019). Verkkoharavoinnissa hyödynnetään tiedonlouhinta-algoritmeja, jotka erottelevat annetun syötteen perusteella kiinnostavan informaation verkkosivuston sisällöstä. Syöte voi koostua esimerkiksi itsenäisiin kokonaisuuksiin eli entiteetteihin liittyvistä nimistä, kuvista, päivämäärä- tai sijaintitiedoista. OSINTissa käytettävät tiedonkeräysmenetelmät voidaan Ünverin (2018) mukaan jakaa karkeasti neljään kategoriaan: tekstitulkinta-, paikkatieto-, kuvatulkinta- sekä verkostoanalyysimenetelmiin. Tässä alaluvussa tarkastellaan edellä mai-

nittuja tiedonkeräysmenetelmiä sekä niiden hyödyntämismahdollisuuksia tiedustelussa.

3.2.1 Tekstitulkintamenetelmät

Tekstin ja kielen tulkintaan perustuvat tiedonkeräysmenetelmät keskittyvät lähteen tekstisisällön jäsentämiseen ja tulkitsemiseen (Gibson 2016, 83). Tekstitulkintaohjelmat erikoistuvat esimerkiksi tekstin semanttisen sisällön tulkitsemiseen, kirjoittajan asenteiden ja arvojen tulkitsemiseen tai tietyn informaation etsimiseen avainsanojen perusteella (Pastor-Galindo ym. 2020; Ünver 2018).

Semanttiset tulkintamenetelmät keskittyvät arvioimaan tekstin välittämiä merkityksiä sekä luokittelemaan tekstiä aiheen, tekstilajin ja tekstissä välitettävien tunteiden perusteella (Pastor-Galindo ym. 2020). Tekstin semanttisen sisällön arvioimiseksi tekstiä jäsennetään sanojen, lauseiden ja kappaleiden tasolla. Sanavalintoja arvioimalla voidaan päätellä tekstin aihepiiri ja kirjoittajan asenne tekstissä käsiteltäviä asioita kohtaan (Gibson, Ramwell ja Day 2016, 98; Ünver 2018). Lauseissa esiintyvien entiteettien välisiä suhteita voidaan arvioida tulkitsemalla niiden kieliopillista asemaa lauseenjäseninä (Best 2011).

Tietyn informaation etsimisessä hyödynnetään erityisesti nimettyjen kokonaisuuksien tunnistusmenetelmiä (engl. *Named-Entity Recognition, NER*). Nimettyjä kokonaisuuksia ovat kaikki entiteetit, jotka voidaan yksilöidä ja erottaa muista saman kategorian entiteeteistä (Di Pietro ym. 2014). Esimerkiksi henkilöt, esineet ja sijainnit ovat kaikki nimettyjä kokonaisuuksia. Nimettyjen kokonaisuuksien tunnistusmenetelmät perustuvat avainsanojen tunnistamiseen suoraan tekstistä ja lauseiden jäsentämiseen sanaluokkien ja kieliopin perusteella (Gibson 2016, 86).

Tekstitulkintamenetelmät ovat käytettävyydeltään joustavia hyvin erilaisten tiedontarpeiden tyydyttämisessä. Tekstimuodossa julkaistava informaatio on asiasisällöltään tiivistä ja kykenee välittämään tietoa myös abstrakteista ilmiöistä, kuten henkilötiedoista, organisaatiorakenteista ja ajatuksista. Raakaa tekstitietoa on helppo ladata ja käsitellä nopeasti, koska se vie suhteellisen vähän tallennustilaa verrattuna audio-, kuva- tai videomateriaaliin.

Tekstitulkintamenetelmien suurimpana heikkoutena ovat kieleen liittyvät haasteet. Puhekieliset ilmaukset, kirjoitusvirheet ja toisesta kielestä käännetyt ilmaukset vaikeuttavat teksti-

tulkintaa (Gibson 2016, 83; Rahwan 2022). Tämän lisäksi lähteen kieli voi olla tekstitulkintaohjelmalle vierasperäistä, mikä rajoittaa tekstin jäsentämistä ja sisällön semanttista luokittelua (Di Pietro ym. 2014). Kieleen liittyvien haasteiden lisäksi tekstimuotoinen sisältö on lähes aina tarkoituksellisesti kirjoitettua ja rajattua, jolloin sisältöön ei usein lipsahda arkaluontoista tietoa, jota tekstin julkaisija ei halua levitettävän internetissä.

3.2.2 Paikkatietomenetelmät

Paikkatieto on kohteeseen tai ilmiöön liittyvää informaatiota, joka viittaa tiettyyn maantieteelliseen sijaintiin. Paikkatieto voi olla luonteeltaan staattista tai dynaamista informaatiota. Staattista paikkatietoa ovat yksittäiseen muuttumattomaan sijaintiin liittyvät tiedot, kuten videon kuvauspaikka tai rakennuksen sijainti kartalla (Stock ja Guesgen 2016, 171). Dynaamisella paikkatiedolla tarkoitetaan sijaintia muuttavaan kohteeseen liitettyä tietoa (Stock ja Guesgen 2016, 171). Tällaista tietoa ovat esimerkiksi lentokoneen sijainnin ja korkeuden ilmaiseva paikkatieto.

Geopaikannuksella tarkoitetaan paikkatietojen selvittämistä hallussa olevien lähtötietojen perusteella (Best 2011). Geopaikannukseen liittyy läheisesti geokoodaus, joka tarkoittaa sijaintitietojen muuttamista paikannimistä koordinaateiksi ja toisinpäin (Gibson, Ramwell ja Day 2016, 103–104). Geopaikannukseen soveltuvia lähtötietoja ovat esimerkiksi koordinaatit, kuvat ja videot tapahtumapaikasta sekä kirjalliset muistiot, jotka kuvailevat tapahtumapaikkaa. Lähtötietoja analysoimalla on mahdollista tunnistaa maantieteeseen perustuvia vihjeitä, joiden avulla tarkka sijainti voidaan selvittää. Tällaisia vihjeitä ovat esimerkiksi postinumerot, rekisterikilpien maatunnukset, tienviitat sekä IP-osoitteet. Geopaikannusta on käytetty muun muassa sotarikosten selvittämisessä (Bellingcat 2017) ja rikollisten piilopaikkojen paljastamisessa (Van Ess 2019).

Paikkatietojen avulla on mahdollista seurata reaaliajassa kehittyviä tapahtumia, kuten liikenneruuhkia tai mielenosoituksia (Stefanidis, Crooks ja Radzikowski 2013). Paikkatieto voi myös paljastaa arkaluontoista tietoa tiedustelun kohteesta. Esimerkiksi sotilasjoukkojen liikehdintää on seurattu Tinder-käyttäjien sijaintitietojen perusteella (Coakley 2021) ja salaisten sotilastukikohtien sijainteja on paljastunut liikuntasovelluksessa jaetuista harjoituslokeis-

ta (Kozera 2020).

Paikkatietomenetelmien automatisoinnin haasteena voidaan pitää yhdenmukaisten sijaintitietojen heikkoa saatavuutta ja epätarkkuutta. Verkkojulkaisuissa sijaintitietoja jaetaan useimmiten paikanniminä koordinaattien sijaan, mikä vaikeuttaa tarkan sijainnin määrittämistä. Sijaintitietojen epätarkkuuteen vaikuttavat muun muassa samanlaiset maantieteelliset paikannimet, epäviralliset paikannimet ja paikkoihin liittyvien sanallisten kuvauksien monitulkintaisuus (Gibson, Ramwell ja Day 2016, 103–104). Paikkatietojen selvittäminen muiden kuin koordinaattien tai paikannimien avulla vaatii päättelykykyä, johon nykyiset koneoppimismenetelmät eivät vielä yllä, vaikka tutkimukset ovat tuottaneet lupaavia tuloksia (Murgese ym. 2022).

3.2.3 Kuvatulkintamenetelmät

Kuvatulkintamenetelmillä analysoidaan digitaalisten kuvien ja videoiden sisältöä sekä niihin liitettyjä metatietoja. Kuvatulkintamenetelmiä hyödynnetään fyysisen maailman kohteiden tunnistamisessa ja paikantamisessa sekä lähdemateriaalin aitouden todentamisessa (Helmus 2022). Kuvien sisällön tulkinta jaetaan tyypillisesti kahteen pääluokkaan. Kuvantunnistus (engl. *image recognition*) keskittyy erottelemaan ja tunnistamaan kuvassa näkyviä entiteettejä, kuten henkilöitä, esineitä ja ajoneuvoja (Rahwan 2022). Kasvojentunnistus (engl. *facial recognition*) keskittyy kuvassa näkyvien kasvojen erottamiseen, yksilöimiseen ja analysoimiseen (Rahwan 2022). Visuaalisen sisällön lisäksi kuva- ja videotiedostot sisältävät metatietoja, jotka ilmaisevat muun muassa tiedoston pakkaustyypin, kameran ominaisuudet sekä kuvaushetkellä käytetyt asetukset (Ünver 2018). Lisäksi metatiedot voivat sisältää kuvauspaikan maantieteelliset koordinaatit (Toevs 2015).

Kuvatulkinta voi paljastaa tiedustelun kohteesta arkaluontoista tietoa, jota julkaisija ei ole ottanut huomioon julkaistessaan sisältöä verkkoon. Kuvassa näkyvien maastonmuotojen, maamerkkien, säätilan ja varjojen perusteella voidaan esimerkiksi päätellä tarkka kuvanottopaikka ja ajankohta ilman metatietoja (Van Der Weide 2020). Kuvien ja videoiden aitouden arvioimiseksi OSINTissa käytetään koneoppimismenetelmiä ja käänteistä kuvahakua (Helmus 2022). Koneoppimismenetelmillä voidaan automaattisesti havaita poikkeamia kuvamateri-

aalista, jotka voivat kieliä materiaalin peukaloinnista (Helmus 2022). Käänteisen kuvahaun perusteella on mahdollista selvittää kuvan alkuperä, mikä puolestaan voi paljastaa kuvaan liittyviä väärennettyjä sosiaalisen median tilejä tai valeuutisia (McKeown ym. 2014; Toler 2019).

Kuvien ja videoiden automaattinen tulkinta ei ole aivan mutkatonta, vaikka ne tarjoavatkin tiedustelulle paljon informaatiota. Kuvista ja videoista on tullut koneoppimismenetelmiä hyödyntävien deepfake-videoiden myötä yleinen tapa levittää disinformaatiota (Helmus 2022). Väärennettyjen videoiden aitouden vahvistaminen tulee olemaan entistä vaikeampi haaste koneoppimismenetelmien kehittyessä (Helmus 2022). Metatietojen saatavuus ja luotettavuus ovat toinen huolenaihe kuvatulkintamenetelmien käytölle. Monet sosiaalisen median palvelut poistavat kuvista ja videoista metatietoja, jotta niitä ei voisi käyttää hyväksi vahingollisessa toiminnassa (Toevs 2015). Metatietojen aitouden vahvistaminen on lisäksi vaikeampaa kuin kuvien sisällön, sillä tietokoneiden käyttöjärjestelmät eivät yleensä pidä kirjaa metatiedoissa tapahtuvista muutoksista (Toevs 2015).

3.2.4 Verkostoanalyysimenetelmät

Verkostoanalyysimenetelmillä kartoitetaan ja tutkitaan kahden tai useamman entiteetin välisiä suhteita ja niiden muodostamia verkostoja (Ünver 2018). Verkosto voi olla esimerkiksi henkilöistä koostuva ystäväpiiri tai organisaation työntekijöiden muodostama hierarkkinen yhteisö. Ihmisten ja organisaatioiden ohella verkostot voivat kuvastaa elottomien järjestelmien osien suhteita toisiinsa. Esimerkiksi useista erilaisista komponenteista koostuvan koneen toimintaa voidaan arvioida tutkimalla komponenttien välisiä suhteita tai yksittäisen komponentin vaikutusta koneen toimintaan.

Alkujaan verkostoanalyysimenetelmät pohjautuvat matematiikan verkkoteoriaan sekä sosiologiaan (Perez ja Germon 2015; Ball 2016). Verkosto koostuu jäsenistä eli solmuista ja jäsenten välisistä suhteista eli kaarista. Verkosto voi koostua erityyppisistä solmuista ja kaarista, jotka eivät ole samankaltaisia keskenään (Gibson, Ramwell ja Day 2016, 101). Solmut voivat esimerkiksi edustaa henkilöitä ja esineitä samassa verkostossa. Solmujen ja kaarien lisäksi verkostolla on ominaisuuksia, kuten verkottuneisuus eli verkoston tiheys, solmujen vä-

liset etäisyydet sekä solmujen välisten kaarien suuntaisuus ja luonne (Ball 2016). Sosiaalisen median verkostojen ominaisuuksia voidaan arvioida analysoimalla esimerkiksi käyttäjätilien tykkäys- ja kommentointiaktiivisuutta, käyttäjien vuorovaikutuksen kohteita sekä käyttäjien ystävä- ja kontaktilistoja.

Verkostoanalyysillä voidaan hahmottaa verkostoon kuuluvien jäsenten keskinäisiä hierarkioita ja käyttäytymistä (Ünver 2018). Lisäksi verkoston sisäisen dynamiikan tarkastelu voi paljastaa verkoston toiminnan kannalta oleellisia avainhenkilöitä ja prosesseja (Ball 2016). Automatisoidut verkostoanalyysiohjelmat tukevat tiedustelua kartoittamalla toimintaympäristön toimijoiden ja ilmiöiden välisiä yhteyksiä, jotka voivat paljastaa yllättäviäkin asioita verkoston toiminnasta.

4 Tiedonhankintamenetelmien soveltuvuus tiedontarpeisiin

Tässä luvussa arvioidaan edellisessä luvussa esiteltyjen tiedonpaikannus- ja tiedonkeräysmenetelmien soveltuvuutta erilaisiin tiedontarpeisiin. Tarkasteltavia tiedontarpeita ovat henkilöt, organisaatiot, esineet ja tapahtumat. Tarkasteltavien tiedontarpeiden valinta ei perustu aiempaan tieteelliseen tutkimukseen, vaan ne edustavat kohteita, joista tyypillisesti hankitaan OSINT-menetelmillä tietoa. Tiedonhankintamenetelmien soveltuvuuden arvioimiseksi tutkimuksessa tarkastellaan kuhunkin tiedontarpeeseen liittyvien tietojen saatavuutta ja käytettävyyttä tiedontarpeen tyydyttämisessä.

4.1 Henkilöt

Tässä tutkielmassa henkilöllä tarkoitetaan ihmistä, jolla on oma henkilökohtainen identiteetti, jonka perusteella tämä pystytään erottamaan muista ihmisistä. Henkilöön kohdistuvalla OSINTilla pyritään usein selvittämään tämän perustiedot, kuten nimi, yhteystiedot ja ulkonäkö, sekä kartoittamaan tämän sidosryhmiä, kuten perhettä ja ystäväpiiriä (Hassan ja Hijazi 2018, 258). Henkilöön kohdistuva OSINT on eettisesti arka aihe, sillä raja päätöksentekoa tukevan tiedustelun ja vaanimisen välillä voi olla häilyvä.

Henkilöön liittyvien tietojen hankkiminen on riippuvainen tämän digitaalisen jalanjäljen koosta (Ramwell, Day ja Gibson 2016, 198–199). Henkilö, joka julkaisee aktiivisesti sisältöä omalla nimellään esimerkiksi sosiaalisessa mediassa, on luonnollisesti helpompi kohde tiedustelulle kuin henkilö, jolla ei ole julkisia käyttäjätilejä verkkopalveluissa. Erityisesti sosiaalisen median palvelut tarjoavat tiedustelulle arvokasta tietoa henkilön kiinnostuksen kohteista ja sidosryhmistä (Hassan ja Hijazi 2018, 258–259). Monissa verkkopalveluissa käyttäjät voivat kuitenkin itse säädellä käyttäjätilinsä näkyvyyttä ulkopuolisille, mikä vähentää julkisesti saatavilla olevaa tietoa kohteesta (McKeown ym. 2014).

Henkilöihin liittyvien tietojen epävarmuus on tunnistettu OSINTissa merkittäväksi haasteeksi, sillä kaikki henkilöstä saatavilla oleva tieto ei välttämättä täsmää tiedustelun todellisen

kohdehenkilön tietojen kanssa (McKeown ym. 2014). Riittämättömät lähtötiedot voivat sekavoittaa tutkimusta varsinkin, jos tietoja löytyy useasta samannimisestä henkilöstä (Weir 2016). Lisäksi henkilöihin liittyvien tietojen vahvistamista vaikeuttavat keksittyjen identiteettien turvin verkossa esiintyvät henkilöt (Ranaldi ja Zanzotto 2020).

Tiedonpaikannusmenetelmistä ohjelmointirajapinnat ja hakurobotit soveltuvat parhaiten henkilöön liittyvien tietojen selvittämiseksi. Ohjelmointirajapinnat mahdollistavat yksittäisten käyttäjätilien ja julkaisujen etsimisen samanaikaisesti useista sosiaalisen median palveluista (Gibson 2016, 76–77; Silvestri ym. 2015). Hakuroboteilla voidaan kartoittaa käyttäjän julkaisuja ja seuraajalistoja (Aliprandi ym. 2014). Verkkosyötteet voivat olla hyödyllisiä tiedonpaikannusmenetelmiä, mikäli tiedustelun kohteena oleva henkilö esiintyy toistuvasti uutisotsikoissa tai julkaisee aktiivisesti sisältöä verkkosyötteitä käyttävissä palveluissa.

Automaattisilla tekstitulkintamenetelmillä on helppo kerätä henkilötietoja sosiaalisen median palveluista, koska käyttäjien profiilisivut ovat yleensä rakenteeltaan yhteneviä samassa palvelussa. Silvestrin ym. (2015) kehittämällä työkalulla käyttäjätileistä kerättyjen nimimerkkien ja yhteystietojen avulla voitiin tunnistaa saman henkilön käyttäjätilejä muista verkkopalveluista automaattisesti. Henkilötietojen ohella tekstitulkintamenetelmillä voidaan arvioida kohdehenkilön asenteita ja arvomaailmaa tulkitsemalla tämän julkaisujen tekstisisältöä ja kommentteja (Drus ja Khalid 2019).

Automaattisilla paikkatietomenetelmillä on mahdollista ennustaa henkilön ajankohtainen olinpaikka tai todennäköinen asuinpaikka (Luo ym. 2020). Esimerkiksi sosiaalisen median julkaisuun liitetyn paikkamerkinnän perusteella voidaan päätellä henkilön sijainti kuvanottohetkellä tai julkaisuhetkellä. Luo ym. (2020) toteavat henkilöiden reaaliaikaisen sijainnin selvittämisen nykyisin käytössä olevilla menetelmillä haasteelliseksi.

Automaattisilla kuvatulkintamenetelmillä voidaan Ranaldin ja Zanzotton (2020) havaintojen mukaan tunnistaa yksittäisen henkilön muita käyttäjätilejä sosiaalisen median julkaisuihin liitettyjen kuvien, symbolien ja logojen perusteella. Erityisesti kasvojentunnistus on havaittu tehokkaaksi keinoksi yksittäisen henkilön eri käyttäjätilien yhdistämisessä (Ranaldi ja Zanzotto 2020).

Automaattisilla verkostoanalyysimenetelmillä voidaan arvioida henkilön todennäköinen a-

suin- tai työpaikka kartoittamalla tämän sosiaalisen median ystäväpiiriä (Luo ym. 2020). Jurgensin (2021) tutkimuksen perusteella henkilön asuinpaikka voidaan päätellä hyvinkin tarkasti analysoimalla tämän sosiaalisen median kontakteja. Henkilön työ- ja harrastusyhteisöjä voidaan yhtä lailla päätellä sosiaalisen median julkaisujen ja kontaktien perusteella.

4.2 Organisaatiot

Organisaatiot käsitetään tässä tutkielmassa yhdestä tai useammasta ihmisestä koostuvaksi ryhmäksi, joka voidaan erottaa toiminnan tavoitteiden ja jäsenten perusteella muista ryhmistä. Yksittäinen ihminen voidaan käsittää henkilön sijaan ryhmäksi, jos kyseessä on esimerkiksi yritys tai järjestö, jonka toiminta voidaan erottaa henkilön yksityiselämästä. Organisaatiot ovat olemukseltaan dynaamisia järjestelmiä, mikä tarkoittaa, että niihin liittyy uusia jäseniä ja niistä poistuu vanhoja jäseniä ajan kuluessa (Bartal, Sasson ja Ravid 2009). Organisaatioiden tiedustelussa mielenkiinto kohdistuu usein yksittäisen organisaation aikeiden, toimintaprotokollien, resurssien ja toimintakyvyn selvittämiseen. Henkilöiden tavoin organisaatioista saatavan tiedon määrä riippuu organisaation julkisuusasteesta. Kaupalliset yritykset ovat lähtökohtaisesti helpompia kohteita tiedustelulle kuin esimerkiksi harrastelijapiirit tai järjestäytyneen rikollisuuden organisaatiot.

Tiedonpaikannusmenetelmistä organisaatioihin liittyviä tietoja on tehokkainta selvittää hakurobottien avulla. Julkisilla organisaatioilla on yleensä verkkosivusto ja sosiaalisen median tilejä, joiden kautta viestitään organisaation ulkopuolisille henkilöille ja sidosryhmille. Tämän lisäksi organisaatioilla saattaa olla sosiaalisen median yhteisöjä, joissa organisaation jäsenet viestivät keskenään sisäisesti. Ei-julkisten organisaatioiden kartoittaminen onnistuu myös hakuroboteilla esimerkiksi käymällä läpi organisaation oletettujen jäsenten sosiaalisen median kontaktilistoja (Aliprandi ym. 2014). Verkkosyötteillä voidaan seurata organisaation ulkoista viestintää ja ilmaantuvuutta mediassa. Ohjelmointirajapinnoilla puolestaan voidaan hankkia tietoja julkisen organisaation tietokannoista organisaation asettamien rajoitteiden puitteissa.

Verkoston sisäistä ja ulkoista viestintää tulkitsemalla voidaan kartoittaa organisaation rakennetta, jäsenten välistä hierarkiaa sekä sidosryhmiä. Automaattiset verkostoanalyysi- ja

tekstitulkintamenetelmät soveltuvat hyvin tällaiseen tarkoitukseen (Ball 2016). Tekstitulkintamenetelmillä voidaan tarkastella organisaation jäsenten viestien asiasisältöä, mistä voi olla hyötyä esimerkiksi organisaation jäsenten aseman ja jäsenten välisten suhteiden laadun arvioimisessa (Ball 2016). Tekstitulkintamenetelmillä voidaan myös arvioida organisaation ulkoista viestintää kuluttajille, yhteistyökumppaneille ja muille sidosryhmille.

Verkostoanalyysimenetelmiä voidaan yksittäiseen henkilöön kohdistuvan tiedustelun tavoin hyödyntää useamman henkilön muodostaman verkoston kartoittamisessa. Verkoston jäsenten viestintäkäyttäytymistä analysoimalla voidaan arvioida suhteiden tiiveyttä ja luonnetta (Chaudhary ja Bansal 2022). Lubarskin ja Morzyn (2012) havaintojen mukaan verkoston jäsen koettiin sitä tärkeämmäksi, mitä nopeammin muut jäsenet reagoivat tämän lähettämiin sähköpostiviesteihin. Havainto voidaan yleistää koskemaan kaikkea henkilöiden välistä viestintää, jolloin esimerkiksi sosiaalisessa mediassa käydyt julkiset keskustelut ja kommentit voivat tarjota arvokasta tietoa jäsenten välisistä suhteista.

Paikkatieto- ja kuvatulkintamenetelmät soveltuvat verrattain heikosti tiedonhankintaan organisaatioista. Organisaatioiden mahdollisia toimipaikkoja voidaan selvittää esimerkiksi julkaisujen sijaintien perusteella (Luo ym. 2020). Kuvatulkintamenetelmiä voidaan puolestaan hyödyntää organisaatioon liittyvien tunnuksien, logojen ja symbolien tunnistamisessa (Ranaldi ja Zanzotto 2020). Kuvatulkintamenetelmistä voi myös olla hyötyä organisaation toimipaikkojen tai vaikutusalueiden tunnistamiseksi kuvamateriaalista. Lisäksi kasvojentunnistusta voidaan hyödyntää organisaation jäsenten kartoittamisessa esimerkiksi yhteiskuvista.

4.3 Esineet

Tässä tutkielmassa esineillä tarkoitetaan elottomia, konkreettisia entiteettejä, jotka voidaan erottaa muista saman kategorian edustajista. Esimerkiksi rakennukset, ajoneuvot ja laitteet luokitellaan esineiksi. Henkilöistä, organisaatioista ja tapahtumista poiketen esineistä ei ole aina mielekästä kerätä tietoa kappalekohtaisesti.

Tiedonpaikannusmenetelmistä hakurobotit soveltuvat parhaiten tietojen keräämiseen esineistä. Esimerkiksi esineistä koostuvia tietokantoja voidaan kartoittaa hakurobottien avulla ja avainsanoja hyödyntävillä hakuroboteilla voidaan etsiä tietoa tietystä esineen mallista. Oh-

jelmointirajapintoja voidaan käyttää esimerkiksi tietosanakirjoista ja katalogeista saatavien tietojen keräämisessä. Verkkosyötteet soveltuvat heikosti tietojen hankkimiseen esineistä, sillä yleensä esineistä julkaistaan tietoa blogeissa ja uutisissa vain harvoin.

Automaattisilla tekstitulkintamenetelmillä voidaan kartoittaa perustietoja esineiden ominaisuuksista, kuten koosta, valmistustiedoista, tuotantomääristä ja käyttötarkoituksesta.

Automaattiset paikkatietomenetelmät soveltuvat suurten kiinteiden esineiden, kuten rakennusten paikantamiseen. Automaattisten kuvatulkintamenetelmien soveltaminen yhdessä paikkatietomenetelmien kanssa voi nopeuttaa huomattavasti rakennusten sijainnin selvittämistä (Murgese ym. 2022). Kiinteiden kohteiden paikantamisen lisäksi liikkuvien ajoneuvojen, kuten laivojen ja lentokoneiden liikettä voidaan seurata niiden lähettämien sijaintitietojen perusteella (Fiorella 2019). Ajoneuvoja pienempien esineiden paikantaminen ei yleensä ole mielekästä eikä usein edes mahdollista.

Esineiden fyysisten ominaisuuksien arvioimisessa kuvatulkintamenetelmät soveltuvat erinomaisesti esineiden tunnistamiseen sekä kappalekohtaisesti että yleisluontoisesti. Sasin, Nairin ja P:n (2022) kehittämää kuvatulkintaohjelmaa on käytetty laittomien aineiden tunnistamiseksi pimeän verkon kauppapaikoilla. Lee ja Kang (2019) puolestaan havaitsivat automaattisen kuvatulkinnan toimivaksi menetelmäksi automerkkien ja mallien tunnistamisessa. Uniikkien esineiden tunnistamiseksi kuvatulkintamenetelmiä voisi soveltaa esimerkiksi auton rekisterikilpien automaattisessa tunnistamisessa.

Viitteitä verkostoanalyysimenetelmien soveltamisesta esineiden välisten verkostojen tulkintaan ei löytynyt tutkimustietoa. Voidaan olettaa, etteivät verkostoanalyysimenetelmät sovellu kovin hyvin esineistä saatavien tietojen keräämiseen, koska esineet eivät julkaise tietoa itsestään.

4.4 Tapahtumat

Tapahtuma käsitetään tässä tutkielmassa ainutkertaisena tai toistuvana ilmiönä, joka tapahtuu tiettynä ajanhetkenä. Tapahtumista saatavilla olevien tietojen määrä vaihtelee merkittävästi tapahtuman julkisuusasteen mukaan. Tapahtumat voivat olla useamman päivän kestä-

viä, monivaiheisia ilmiöitä, joista on saatavilla paljon erilaista tietoa. Tällaisia tapahtumia ovat esimerkiksi luonnonilmiöt, festivaalit tai taistelut. Toisaalta tapahtumat voivat olla hyvin lyhytkestoisia tapahtumia, joista ei välttämättä ole saatavilla tietoa etukäteen eikä tapahtuman jälkeen. Esimerkiksi kahden henkilön välinen tapaaminen voi olla julkisuusasteeltaan matala, jolloin siitä ei todennäköisesti ilmesty tietoa julkisesti saataville.

Tapahtumien ajallisen ulottuvuuden takia ohjelmointirajapinnat ja verkkosyötteet toimivat parhaiten reaaliaikaisten tapahtumien seuraamisessa ja tulevaisuuden tapahtumiin liittyvän informaation hankkimisessa. Verkkosyötteet voivat olla erityisen hyödyllisiä kehittyvien tapahtumien seurannassa, mikäli ne ylittävät uutiskynnyksen (European Media Monitor 2023). Ohjelmointirajapinnoilla on mahdollista kerätä tietoa erityisillä hashtageilla tai muilla tunnisteilla tehdyistä julkaisusta reaaliaikaisesti (Stefanidis, Crooks ja Radzikowski 2013). Tunnisteiden ilmaantuvuuden seuraaminen sosiaalisen median kanavissa ja uutispalveluissa voi auttaa tiedustelua arvioimaan tapahtuman merkittävyyttä (Best 2011). Hakuroboteilla on teoriassa mahdollista hankkia tietoa myös käynnissä olevista tapahtumista, mutta niiden toimintafilosofia soveltuu paremmin laajojen tietokantojen kartoittamiseen.

Tekstitulkintamenetelmillä voidaan kerätä tapahtumaan liittyviä perustietoja, kuten sijainti, ajankohta ja järjestäjä (Li, Ji ja Zhao 2015). Tämän lisäksi tekstitulkintamenetelmillä voidaan kartoittaa tapahtumaan osallistuvia tahoja (Li, Ji ja Zhao 2015). Semanttisilla tekstintulkintamenetelmillä kyetään luokittelemaan erillisiä tapahtumia niistä saatavilla olevia tietoja vertailemalla eri lähteiden välillä (Yang ja Lee 2012). Zhang ym. (2017) havaitsivat tapahtumasta kertovien tekstisisältöjen tulkintatarkkuuden parantuvan yhdistämällä tekstitulkintamenetelmien lisäksi kuvatulkintamenetelmiä analyysiin. Liittämällä tekstitulkintatyökalun jäsentämiin entiteetteihin kuvia entiteettien nimen perusteella analyysin tarkkuus parani huomattavasti tekstin semanttisen tulkinnan osalta (Zhang ym. 2017).

Kuvatulkintamenetelmät ovat tapahtuman olosuhteiden ymmärtämiseksi korvaamaton työkalu, sillä kuvat välittävät tietoa tapahtumaan osallistuvista henkilöistä ja tapahtuman vaiheista, joista ei välttämättä löydy kirjoitettua tietoa. Kuvat tapahtumapaikalta voivat paljastaa tapahtuman tarkan maantieteellisen sijainnin ja sen myötä kuvanottoajankohdan (Van Der Weide 2020). Murgesen ym. (2022) havaintojen perusteella kuvia on mahdollista geopaikantaa automaattisesti alueilla, joista on olemassa valmiiksi paljon kuvamateriaalia. Ta-

pahtumien geopaikantamisesta olisi suuri hyöty onnettomuustilanteissa, joissa apu täytyy toimittaa nopeasti perille (Murgese ym. 2022).

Tapahtumaan liittyviä paikkatietoja on mahdollista hankkia esimerkiksi sosiaalisen median julkaisuista (Stefanidis, Crooks ja Radzikowski 2013). Julkaisuihin liittyvien sijaintitietojen perusteella voidaan arvioida tapahtuman kokoa ja tapahtuman sijainnin muutosta reaaliaikaisesti (Stefanidis, Crooks ja Radzikowski 2013). Koska tapahtumaan liittyy useita erityyppisiä entiteettejä, voidaan verkostoanalyysimenetelmillä tulkita eri henkilöiden, ryhmien ja esineiden osallisuutta tapahtumaan.

5 Yhteenveto

Havaintojen perusteella tiedonpaikannusmenetelmistä soveltuvimmaksi arvioidaan hakurobotit. Tiedonpaikannusmenetelmien soveltuvuuteen vaikuttavat ratkaisevasti tarvittavan tiedon ajankohtaisuus ja julkisuusaste. Verkkosyötteet ja ohjelmointirajapinnat ovat toimivia menetelmiä reaaliaikaisten tietojen hankkimisessa, mutta ne soveltuvat heikosti matalan julkisuusasteen tietojen paikantamiseen. Hakurobotit puolestaan soveltuvat perustietojen keräämiseen suurista tietomassoista, joista on olemassa pidemmältä aikaväliltä dokumentoitua tietoa. Hakuroboteilla kyetään hankkimaan tietoa myös matalan julkisuusasteen kohteista.

Tiedonkeräysmenetelmistä teksti- ja kuvatulkintamenetelmät arvioidaan monikäyttöisimmiksi, sillä tekstit ja kuvat välittävät paljon tietoa tiiviisti ilmaistuna erilaisista kohteista. Tekstitulkintamenetelmillä kyetään etsimään tietoa konkreettisten ja abstraktien kohteiden tuntomerkkien perusteella. Kuvatulkintamenetelmillä voidaan analysoida konkreettisten kohteiden ominaisuuksia ja kuvan olosuhteita päätelmien tekemiseksi. Paikkatieto- ja verkostoanalyysimenetelmät soveltuvat heikoiten erilaisten tiedontarpeiden tyydyttämisessä. Paikkatietomenetelmillä voidaan lähinnä hankkia tietoa tiedontarpeen fyysisestä sijainnista ja verkostoanalyysimenetelmät soveltuvat kohteiden vuorovaikutuksen arviointiin. Henkilöiden ja organisaatioiden välisen vuorovaikutuksen arviointi voi paljastaa niiden toimintaa ohjaavia aikeita.

Tiedonhankinnan automaation keskeisiä haasteita ovat tiedon määrään, laatuun, relevanssiin ja luotettavuuteen liittyvät ongelmat, jotka tulisi ottaa huomioon tiedonhankintaohjelmia suunniteltaessa. Tiedustelun keräysvaiheessa käsiteltävän tiedon määrän vähentämiseksi on tärkeää, että tiedonhankintaohjelmalle annettu syöte on riittävän yksityiskohtaista ja että ohjelma kykenee käsittelemään yksityiskohtaisia lähtötietoja hakua rajoittavina parametreina. Tiedonhankintaohjelman tulisi kyetä tunnistamaan toisteista informaatiota jäsenninalgoritmien avulla, jotta käsitys havaitun ilmiön merkittävyydestä ei vinoutuisi. Laadukkaiksi havaitut lähteet ja niiden sisältämä tieto tulisi tallentaa ja indeksoida myöhempää käyttöä varten, jotta tiedonhankintaa ei tarvitsisi suorittaa tarpeettomasti uudestaan. Kerätyn tiedon oikeellisuuden arvioimiseksi tiedonhankintaohjelmien tulisi arvioida lähdettä ja sen sisältämää informaatiota vertailemalla sitä muihin samasta ilmiöistä kirjoitettuihin julkaisuihin

sekä pitämällä kirjaa lähteen aikaisemmin julkaiseman informaation oikeellisuudesta.

Lopuksi voidaan todeta, että avointen lähteiden tiedustelun tiedonhankintamenetelmien automaattisia sovellusmahdollisuuksia tulisi tutkia aiempaa enemmän. Tulevaisuudessa tietojen analysointi ja tiedolla johtamisen taito tulevat todennäköisesti korostumaan päätöksenteossa. Alati kasvava julkisen tiedon määrä tulee muodostamaan haasteita toimintaympäristön ennakoimiselle, jolloin tiedustelun on ensiarvoisen tärkeää kyetä erottelemaan tehokkaasti oleellista, tarkkaa ja luotettavaa informaatiota tietovirrasta.

Lähteet

Aliprandi, Carlo, Juan A. Irujo, Montse Cuadros, Sebastian Maier, Felipe Melero ja Matteo Raffaelli. 2014. “CAPER: Collaborative Information, Acquisition, Processing, Exploitation and Reporting for the Prevention of Organised Crime”. Teoksessa *HCI International 2014 - Posters’ Extended Abstracts. Communications in Computer and Information Science,* 434:147–152. Cham, Sveitsi: Springer. https://doi.org/10.1007/978-3-319-07857-1_26.

Ball, Leslie. 2016. “Automating social network analysis: A power tool for counter-terrorism”. *Security Journal* 29:147–168. https://doi.org/10.1057/sj.2013.3.

Bartal, Alon, Elan Sasson ja Gilad Ravid. 2009. “Predicting Links in Social Networks using Text Mining and SNA”. Teoksessa *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining,* 131–136. IEEE. https://doi.org/10.1109/ASONAM.2009.12.

Bayerl, Petra S., Babak Akhgar, Alice Raven, Helen Gibson ja Tony Day. 2023. “Future Challenges and Requirements for Open Source Intelligence in Law Enforcement Investigations: Results from a horizon scanning exercise”. *European Law Enforcement Research Bulletin* 22 (22): 21–38. https://bulletin.cepol.europa.eu/index.php/bulletin/issue/view/31.

Bazzell, Michael. 2019. *Open source intelligence techniques: Resources for searching and analyzing online information.* 7. painos. Itsenäisesti julkaistu. https://inteltechniques.com.

Bellingcat. 2017. *How a Werfalli Execution Site Was Geolocated.* https://www.bellingcat.com/news/mena/2017/10/03/how-an-execution-site-was-geolocated/. Viitattu 11.4.2023.

———. 2022. *Bellingcat Online Investigation Toolkit.* http://bit.ly/bcattools. Viitattu 18.3.2023.

Benes, Libor. 2013. “OSINT, New Technologies, Education: Expanding Opportunities and Threats. A New Paradigm”. *Journal of Strategic Security* 6 (3): 22–37. https://www.jstor.org/stable/26485053.

Best, Clive. 2008. “Web Mining for Open Source Intelligence”. Teoksessa *2008 12th International Conference Information Visualisation,* 321–325. IEEE. https://doi.org/10.1109/IV.2008.86.

———. 2011. “Challenges in Open Source Intelligence”. Teoksessa *2011 European Intelligence and Security Informatics Conference,* 58–62. IEEE. https://doi.org/10.1109/EISIC.2011.41.

Chaudhary, Megan, ja Divya Bansal. 2022. “Open source intelligence extraction for terrorism-related information: A review”. *WIREs Data Mining and Knowledge Discovery* 12 (5). https://doi.org/10.1002/widm.1473.

Coakley, Amanda. 2021. “Borderline: Tinder profiles of Polish troops appear in Belarus”. *The Independent*, 15.11.2021, https://www.independent.co.uk/news/world/europe/belarus-poland-border-tinder-troops-b1957953.html.

Derr, Richard L. 1983. “A conceptual analysis of information need”. *Information Processing & Management* 19 (5): 273–278. https://doi.org/10.1016/0306-4573(83)90001-8.

Di Pietro, Giulia, Carlo Aliprandi, Antonio E. De Luca, Matteo Raffaelli ja Tiziana Soru. 2014. “Semantic crawling: An approach based on Named Entity Recognition”. Teoksessa *2014 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining, ASONAM,* 695–699. IEEE. https://doi.org/10.1109/ASONAM.2014.6921661.

Diouf, Rabiyatou, Edouard Ngor Sarr, Ousmane Sall, Babiga Birregah, Mamadou Bousso ja Sény Ndiaye Mbaye. 2019. “Web Scraping: State-of-the-Art and Areas of Application”. Teoksessa *2019 IEEE International Conference on Big Data (Big Data),* 6040–6042. IEEE. https://doi.org/10.1109/BigData47090.2019.9005594.

Drus, Zulfadzli, ja Haliyana Khalid. 2019. “Sentiment Analysis in Social Media and Its Application: Systematic Literature Review”. *Procedia Computer Science* 161:707–714. https://doi.org/10.1016/j.procs.2019.11.174.

Dupont, Alan. 2003. “Intelligence for the Twenty-First Century”. *Intelligence and National Security* 18 (4): 15–39. https://doi.org/10.1080/02684520310001688862.

Eldridge, Christopher, Christopher Hobbs ja Matthew Moran. 2017. “Fusing algorithms and analysts: open-source intelligence in the age of ‘Big Data’”. *Intelligence and National Security* 33 (3): 391–406. https://doi.org/10.1080/02684527.2017.1406677.

European Media Monitor, EU. 2023. *EMM Current top 10 stories.* https://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html. Viitattu 10.4.2023.

Fiorella, Giancarlo. 2019. *A Beginner’s Guide To Flight Tracking.* https://www.bellingcat.com/resources/how-tos/2019/10/15/a-beginners-guide-to-flight-tracking/. Viitattu 29.4.2023.

Gibson, Helen. 2016. “Acquisition and Preparation of Data for OSINT Investigations”. Teoksessa *Open Source Intelligence Investigation From Strategy to Implementation,* toimittanut Babak Akhgar, Saskia M. Bayerl ja Fraser Sampson, 69–94. Springer International Publishing. https://doi.org/10.1007/978-3-319-47671-1.

Gibson, Helen, Steve Ramwell ja Tony Day. 2016. “Analysis, Interpretation and Validation of Open Source Data”. Teoksessa *Open Source Intelligence Investigation From Strategy to Implementation,* toimittanut Babak Akhgar, Saskia M. Bayerl ja Fraser Sampson, 95–110. Springer International Publishing. https://doi.org/10.1007/978-3-319-47671-1.

Haasio, Ari, ja Reijo Savolainen. 2004. *Tiedonhankintatutkimuksen perusteet.* Helsinki: BTJ.

Hassan, Nihad A., ja Rami Hijazi. 2018. *Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence.* Berkeley, Kalifornia: Apress. https://doi.org/10.1007/978-1-4842-3213-2.

Helmus, Todd C. 2022. “Artificial Intelligence, Deepfakes, and Disinformation: A Primer”. *RAND Corporation,* http://www.jstor.org/stable/resrep42027.

Hoppa, Mary A., Scott M. Debb, George Hsieh ja KC Bigyan. 2020. “TwitterOSINT: Automated Open Source Intelligence Collection, Analysis & Visualization”. Teoksessa *Annual Review of Cybertherapy and Telemedicine 2019,* toimittanut Brenda K. Wiederhold, Giuseppe Riva ja Scott M. Debb, 17:121–128. Interactive Media Institute.

Hulnick, Arthur S. 2002. “The Downside of Open Source Intelligence”. *International Journal of Intelligence and Counterintelligence* 15 (4): 565–579. https://doi.org/10.1080/08850600290101767.

Jardines, Eliot A. 2016. “Open Source Intelligence”. Teoksessa *The Five Disciplines of Intelligence Collection,* toimittanut Mark M. Lowenthal ja Robert M. Clark, 5–43. Thousand Oaks, Kalifornia: CQ Press.

JCS, Yhdysvaltain puolustushaarakomentajien neuvosto. 2019. *Insights and Best Practices Focus Paper: Intelligence Operations.* https://www.jcs.mil/Doctrine/focus_papers/. Viitattu 11.4.2023. Haettu 11.4.2023 osoitteesta: https://www.jcs.mil/Portals/36/Documents/Doctrine/ fp/intell_ops_fp.pdf.

———. 2021. *DOD Dictionary of Military and Associated Terms, November 2021.* http://www.jcs.mil/Doctrine/DOD-Terminology/. Viitattu 9.3.2023. Haettu 9.3.2023 osoitteesta: https://irp.fas.org/doddir/dod/dictionary.pdf.

Jurgens, David. 2021. “That’s What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships”. Teoksessa *Proceedings of the International AAAI Conference on Web and Social Media,* 7:273–282. 1. https://doi.org/10.1609/icwsm.v7i1.14399.

Järvelin, Kalervo, ja Eero Sormunen. 2010. “Tiedon tallennus ja haku”. Teoksessa *Ote informaatiosta: Johdatus informaatiotutkimukseen ja interaktiiviseen mediaan,* toimittanut Sami Serola, 155–210. Helsinki: BTJ.

Kozera, Cyprian A. 2020. “Fitness OSINT: Identifying and tracking military and security personnel with fitness applications for intelligence gathering purposes”. *Security and Defence Quarterly* (Varsova) 32 (5): 41–52. https://doi.org/10.35467/sdq/131759.

Lee, Yunsoo, ja Suk-Ju Kang. 2019. “Web Scraping Crawling-based Automatic Data Augmentation for Deep Neural Networks-based Vehicle Classifications”. Teoksessa *2019 IEEE International Conference on Consumer Electronics (ICCE),* 1–2. Las Vegas, Nevada: IEEE. https://doi.org/10.1109/ICCE.2019.8661971.

Li, Hao, Heng Ji ja Lin Zhao. 2015. “Social Event Extraction: Task, Challenges and Techniques”. Teoksessa *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* 526–532. Association for Computing Machinery. https://doi.org/10.1145/2808797.2809413.

Lowenthal, Mark M. 2020. *Intelligence: From Secrets to Policy.* 8. painos. Thousand Oaks, Kalifornia: CQ Press.

Lubarski, Pawel, ja Mikolaj Morzy. 2012. “Measuring the Importance of Users in a Social Network Based on Email Communication Patterns”. Teoksessa *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* 86–90. IEEE. https://doi.org/10.1109/ASONAM.2012.24.

Luo, Xiangyang, Yaqiong Qiao, Chenliang Li, Jiangtao Ma ja Yimin Liu. 2020. “An overview of microblog user geolocation methods”. *Information Processing & Management* 57 (6). https://doi.org/10.1002/widm.1473.

Martelius, Juha. 2020. “Tiedustelutieto kansallisen turvallisuuden päätöksenteossa”. Teoksessa *Suomalaisen tiedustelukulttuurin jäljillä,* toimittanut Tommi Koivula, 57–75. Maanpuolustuskorkeakoulu. https://urn.fi/URN:ISBN:978-951-25-3139-4.

Marzell, Laurence. 2016. “OSINT as Part of the Strategic National Security Landscape”. Teoksessa *Open Source Intelligence Investigation From Strategy to Implementation,* toimittanut Babak Akhgar, Saskia M. Bayerl ja Fraser Sampson, 33–55. Springer International Publishing. https://doi.org/10.1007/978-3-319-47671-1.

McDowell, Don. 2009. *Strategic Intelligence: A Handbook for Practitioners, Managers, and Users.* Lanham, Maryland: Scarecrow Press.

McKeown, Sean, David Maxwell, Leif Azzopardi ja William Bradley Glisson. 2014. “Investigating people: a qualitative analysis of the search behaviours of open-source intelligence analysts”. Teoksessa *Proceedings of the 5th Information Interaction in Context Symposium,* 175–184. New York City, New York: Association for Computing Machinery. https://doi.org/10.1109/GOCICT.2015.14.

Murgese, Fabio, Gerard Alcaina, Oguz Mulayim, Jesus Cerquides ja Jose Luis Fernandez Marquez. 2022. “Automatic Outdoor Image Geolocation with Focal Modulation Networks”. Teoksessa *Frontiers in Artificial Intelligence and Applications,* toimittanut Atia Cortés ja Tommaso Grimaldo Francisco and Flaminio, 356:279–288. IOS Press. https://ebooks.iospress.nl/volumearticle/61255.

NATO. 2002. *NATO Open Source Intelligence Handbook.* https://archive.org/details/NATOOSINTHandbookV1.2/.

Nordine, Justine. 2022. *OSINT Framework.* https://osintframework.com/. Viitattu 18.3.2023.

Pastor-Galindo, Javier, Pantaleone Nespoli, Félix Gómez Mármol ja Gregorio Martínez Pérez. 2020. “The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends”. *IEEExplorer* 8:10282–10304. https://doi.org/10.1109/ACCESS.2020.2965257.

Perez, Charles, ja Rony Germon. 2015. “Graph Creation and Analysis for Linking Actors: Application to Social Data”. Teoksessa *Automating Open Source Intelligence: Algorithms for OSINT,* toimittanut Robert Layton ja Paul A. Watters, 103–129. Elsevier Science & Technology Books. https://doi.org/10.1016/C2014-0-02170-3.

Ponder-Sutton, Agate M. 2016. “The Automating of Open Source Intelligence”. Teoksessa *Automating Open Source Intelligence: Algorithms for OSINT,* toimittanut Robert Layton ja Paul A. Watters, 1–20. Elsevier Science & Technology Books. https://doi.org/10.1016/B978-0-12-802916-9.00001-4.

Porvali, Mikko. 2018. *Tiedustelun Näkymätön Historia: Antiikista Maailmansotiin.* 1. painos. 13. Jyväskylä: Atena.

Rahwan, Amr el. 2022. “Artificial Intelligence and Interoperability for Solving Challenges of OSINT and Cross-Border Investigations”. Teoksessa *European Law Enforcement Research Bulletin,* nide 6. Euroopan Unionin lainvalvontakoulutusvirasto CEPOL. https://bulletin.cepol.europa.eu/index.php/bulletin/article/view/535.

Ramwell, Steve, Tony Day ja Helen Gibson. 2016. “Use Cases and Best Practices for LEAs”. Teoksessa *Open Source Intelligence Investigation From Strategy to Implementation,* toimittanut Babak Akhgar, Saskia M. Bayerl ja Fraser Sampson, 197–211. Springer International Publishing. https://doi.org/10.1007/978-3-319-47671-1.

Ranaldi, Leondardo, ja Fabio M. Zanzotto. 2020. “Hiding Your Face Is Not Enough: user identity linkage with image recognition”. *Social Network Analysis and Mining* 10 (56). https://doi.org/10.1007/s13278-020-00673-4.

Rowley, Jennifer. 2007. “The wisdom hierarchy: representations of the DIKW hierarchy”. *Journal of Information Science* 33 (2): 163–180. https://doi.org/10.1177/0165551506070706.

Sasi, Ashwin, Varun Nair ja Vipin P. 2022. “DARKWEB IMAGE SCRAPPER: An Open Source Intelligence Tool”. Teoksessa *2022 International Conference on Connected Systems & Intelligence (CSI),* 1–6. Trivandrum, Intia: IEEE. https://doi.org/10.1109/CSI54720.2022.9924098.

Savolainen, Reijo. 2010. “Tiedonhankintatutkimuksen lähtökohtia”. Teoksessa *Ote informaatiosta: Johdatus informaatiotutkimukseen ja interaktiiviseen mediaan,* toimittanut Sami Serola, 75–115. Helsinki: BTJ.

Silvestri, Giuseppe, Jie Yang, Alessandro Bozzon ja Andrea Tagarelli. 2015. “Linking Accounts across Social Networks: the Case of StackOverflow, Github and Twitter”. https://api.semanticscholar.org/CorpusID:12483080.

Stefanidis, Anthony, Andrew Crooks ja Jacek Radzikowski. 2013. “Harvesting ambient geospatial information from social media feeds”. *GeoJournal* 78 (2): 319–338. https://www.jstor.org/stable/42006322.

Stock, Kristin, ja Hans Guesgen. 2016. “Geospatial Reasoning With Open Data”. Teoksessa *Automating Open Source Intelligence: Algorithms for OSINT,* toimittanut Robert Layton ja Paul A. Watters, 171–204. Elsevier Science & Technology Books. https://doi.org/10.1016/B978-0-12-802916-9.00001-4.

Toevs, Brian. 2015. “Processing of Metadata on Multimedia Using ExifTool: A Programming Approach in Python”. Teoksessa *2015 Annual Global Online Conference on Information and Computer Technology (GOCICT),* 26–30. IEEE. https://doi.org/10.1109/GOCICT.2015.14.

Toler, Aric. 2019. *Guide To Using Reverse Image Search For Investigations.* https://www.bellingcat.com/resources/how-tos/2019/12/26/guide-to-using-reverse-image-search-for-investigations/. Viitattu 28.4.2023.

Urbansky, David, Sandro Reichert, Klemens Muthmann, Daniel Schuster ja Alexander Schill. 2021. “An Optimized Web Feed Aggregation Approach for Generic Feed Types”. Teoksessa *Proceedings of the International AAAI Conference on Web and Social Media,* 5:638–641. 1. https://doi.org/10.1609/icwsm.v5i1.14161.

Van Der Weide, Youri. 2020. *Using the Sun and the Shadows for Geolocation.* https://www.bellingcat.com/resources/2020/12/03/using-the-sun-and-the-shadows-for-geolocation/. Viitattu 28.4.2023.

Van Ess, Henk. 2019. *Locating The Netherlands’ Most Wanted Criminal By Scrutinising Instagram.* https://www.bellingcat.com/news/uk-and-europe/2019/03/19/locating-the-netherlands-most-wanted-criminal-by-scrutinising-instagram/. Viitattu 28.4.2023.

Warner, Michael. 2002. “Wanted: A Definition of Intelligence”. *Studies in Intelligence* 46 (3): 15–23.

Weir, George R. S. 2016. “The Limitations of Automating OSINT: Understanding the Question, Not the Answer”. Teoksessa *Automating Open Source Intelligence: Algorithms for OSINT,* toimittanut Robert Layton ja Paul A. Watters, 159–169. Elsevier Science & Technology Books. https://doi.org/10.1016/B978-0-12-802916-9.00009-9.

Yang, Hsin-Chang, ja Chung-Hong Lee. 2012. “Mining open source text documents for intelligence gathering”. Teoksessa *2012 International Symposium on Information Technologies in Medicine and Education,* 2:969–973. IEEE. https://doi.org/10.1109/ITiME.2012.6291464.

Ünver, H. Akin. 2018. “Digital Open Source Intelligence and International Security: A Primer”. Centre for Economics ja Foreign Policy Studies. https://www.jstor.org/stable/resrep21048.

Zhang, Tongtao, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji ja Shih-Fu Chang. 2017. “Improving Event Extraction via Multimodal Integration”. Teoksessa *Proceedings of the 25th ACM International Conference on Multimedia,* 270–278. New York City, New York: Association for Computing Machinery. https://doi.org/10.1145/3123266.3123294.