**Tapio Saarnio**

# Hakurobotti ja robots.txt-tiedosto

Tietotekniikan kandidaatintutkielma

30. huhtikuuta 2023

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

**Tekijä:** Tapio Saarnio

**Yhteystiedot:** tapio.j.saarnio@student.jyu.fi

**Ohjaaja:** Tytti Saksa

**Työn nimi:** Hakurobotti ja robots.txt-tiedosto

**Title in English:** Web Crawler and the robots.txt file

**Työ:** Kandidaatintutkielma

**Opintosuunta:** Tietotekniikka

**Sivumäärä:** 20+0

**Tiivistelmä:** Tutkielmassani esitellään kuinka hakurobotti toimii ja mitkä ovat sen keskeisiä käyttötarkoituksia. Lisäksi tutkitaan hakurobottien toiminnan sääntelemiseen kehitettyä tämänhetkistä de facto-standardia: Robots Exclusion -protokollaa, sekä mitä haasteita hakurobottien epäeettisestä toiminnasta voi koitua sekä yksilöille että yhteiskunnalle.

**Avainsanat:** Hakurobotti, Robots.txt, Tiedonharavointi

**Abstract:** In my thesis we focus on the Web Crawler, how does it work and for what purposes are they used for. Additionally the de facto standard in the industry, the Robots Exclusion Protocol is presented. We also discuss what challenges does Web Crawler present to both individuals and for the society as whole.

**Keywords:** Web Crawler, Robots.txt, Web Scraping

# Kuviot

# Sisällys

# 1 Johdanto

World Wide Web voidaan määritellä informaatiota sisältäväksi paikaksi, jossa resursseja voidaan jakaa käyttäjien kesken internetin välityksellä URI (engl. Uniform Resource Identifier) -tunnusten avulla (W3C, n.d.). Web-sivujen lukumäärä, sekä niissä vierailevien ihmisten volyymi ovat kasvaneet ja kasvavat edelleen merkittävästi World Wide Webin keksimisestä (1989) lähtien. Hakukoneet, kuten esimerkiksi Google, Yahoo ja Bing ovat olleet jo 90-luvulta lähtien, ja ovat vielä nykyäänkin yleisin työkalu Webin miljardien sivujen navigoimiseen ja tiedon etsimiseen. Hakukoneet tarvitsevat automatisoituja työkaluja, jotta ne voivat löytää uusia sivuja sekä päivittää olemassa olevia sivuja hakemistoihinsa. Näitä työkaluja kutsutaan hakuroboteiksi (engl. Web Crawler). Viime vuosina datan kasvanut arvo ja uudet innovaatiot ovat johtaneet siihen, että nykyään hakurobotteja käytetään yhä enemmän tietynlaisen datan keräämiseen nettisivuilta ja sen jatkojalostamiseen. Hakurobotit mahdollistavat esimerkiksi monille tutut hintavertailusivustot.

Matthew Gray kehitti ensimmäisen hakurobotin opiskellessaan MIT:ssä vuonna 1993. "The Wanderer"niminen hakurobotti etsi uusia web-sivuja ja keräsi statistiikkaa web-sivujen määrän kasvusta. (Gray, n.d.) Seuraavina vuosina kehitettiin ja julkaistiin lukuisia muita hakurobotteja ja hakukoneita, joihin kuuluvat esimerkiksi WebCrawler, Aliweb, Jumpstation, World Wide Web Worm, Lycos, Altavista sekä Yahoo. Ensimmäiset akateemiset julkaisut hakurobotteihin liittyen julkaistiin ensimmäisen World Wide Web konferenssin yhteydessä vuonna 1994 (Eichmann 1994; Pinkerton 1994; McBryan 1994). Yksi merkittävimmistä akateemisista julkaisuista hakukoneisiin ja hakurobotteihin liittyen on Brin ja Page (1998) artikkeli nimeltään "The anatomy of a large-scalehypertextual Web searchengine", jossa esitellään prototyyppi maailman suosituimmasta hakukoneesta Googlesta.

Tarve rajoittaa hakurobottien toimintaa huomattiin jo varhain niiden keksimisen jälkeen. Aiemmin mainitun maailman ensimmäisenä hakukoneena pidetyn Aliwebin kehittänyt Martijn Koster julkaisi ehdotuksensa Robots Exclusion -protokollasta (REP) vuonna 1994 hakuroboteista kiinnostuneiden kehittäjien sähköpostilistakeskustelujen perusteella (Koster 1994). Kerrotaan, että Koster aloitti protokollan kehittämisen sen jälkeen, kun erään hänen kollegan kehittämä hakurobotti ylikuormitti hänen kotisivunsa toimimattomaksi.

Tässä kandidaatintutkielmassa esittellään hakurobottien yleisiä toimintaperiaatteita, niiden eri käyttötarkoituksia sekä kuinka Robots.txt tiedoston avulla voidaan säädellä hakurobottien toimintaa. Lopuksi käsittellään vielä mitä haittavaikutuksia hakurobotit saattavat pahimmassa tapauksessa aiheuttaa.

# 2 Hakurobotti ja Robots Exclusion -protokolla

## 2.1 Yleistä hakuroboteista

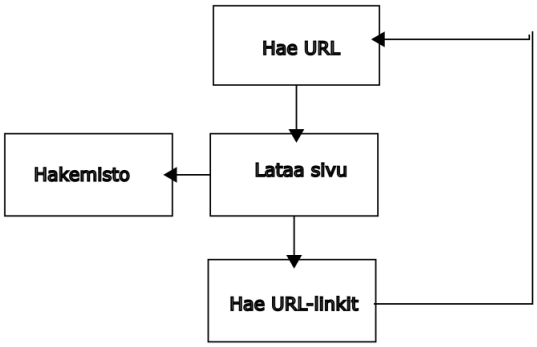
Hakurobotti voidaan määritellä tietokoneohjelmaksi, joka selaa Webiä systemaattisesti ja automaattisesti (Udapure, Kale ja Dharmik 2014). Hakurobotin tärkein tehtävä on kerätä ja tallentaa verkkosivuja lokaaliin hakemistoon (Yu ym. 2020). Tätä prosessia kutsutaan web-sivujen indeksoinniksi. Hakurobottien suorittama indeksointi on tärkeä osa hakukoneiden toimintaa, sillä käyttäjän hakusanoihin täsmäävät palautettavat hakutulokset etsitään hakurobottien keräämästä tietokannasta (Pandey ja Olston 2005). Ei riitä että web-sivu indeksoidaan kerran, vaan hakurobotin on vierailtava samalla sivulla tasaisin väliajoin varmistaakseen, että hakukoneen tajoama tieto pysyy ajankohtaisena (Cho ja Garcia-Molina 2000).

World Wide Webiä voidaan mallintaa suunnattuna graafina, jossa staattiset web-sivut muodostavat graafin solmut ja web-sivujen väliset linkit muodostavat graafin kaaret (Broder ym. 2000). World Wide Webin rakenteen tulkitseminen graafina mahdollistaa graafiteoriasta tuttujen algoritmien soveltamisen hakurobottien käyttämiin hakualgoritmeihin. Tyypillinen hakurobotti liikkuu hyperlinkkejä pitkin sivulta toiseen web-sivujen muodostamaa verkkoa pitkin (Cothey 2004).

Englanninkielisessä kirjallisuudessa sanoja ”Web Spider”, ”Web Wanderer” ja ”Web Robot” käytetään synonyyminä hakurobotin kanssa, joka on käännetty termistä ”Web Crawler”. Toinen aiheeseen läheisesti liittyvä termi on tiedonharavointi (engl. Web Scraping), joka tarkoittaa nettisivujen sisältämän tietynlaisen tiedon etsimistä ja tallentamista hakurobotin avulla. Toisin kuin indeksoinnissa, tiedonharavoinnissa web-sivua ei välttämättä tallenneta kokonaisuudessaan, vaan ainoastaan haluttu tieto.

Kuvio 1 (grafiikka tekijän) havainnollistaa hakurobotin rekursiivista toimintaa. Hakurobotille voidaan antaa aloitus URL, jonka jälkeen se tallentaa sivun hakemistoon ja kerää jokaisen sivun sisältämän linkin talteen. Tämän jälkeen sama prosessi toistuu, aloittaen kerätyistä URL-linkeistä.

Hakurobotit käyttävät erilaisia hakualgoritmeja käyttötarkoituksestaan riippuen. Kumar, Bha-




Kuvio 1. Hakurobotin rekursiivinen toiminta, mukailtu (Kausar, Dhaka ja Singh 2013)

tia ja Rattan (2017) tutkimuksessaan esittelevät 5 kategoriaa joihin erityyppiset hakurobotit voidaan jaotella. Universaali hakurobotti hakee kaikki kohtaamansa web-sivut ja niiden sisältämät URL-linkit karsimatta pois mitään tietoa. Valikoiva hakurobotti kerää ja tallentaa ainoastaan tietyn ehdon täyttävät web-sivut talteen. Ehdoksi voidaan asettaa esimerkiksi tietty aihe tai hakusana. Piilotettu hakurobotti keskittyy syväverkon (engl. Deep Web) tutkimiseen. Mobiili hakurobotti suorittaa web-sivujen filtteröinnin serverin eikä hakukoneen puolella vähentäen verkon kuormitusta. Jatkuvan hakurobotin tehtävänä on päivittää jo löydettyjä sivuja.

## 2.2 Hakurobottien käyttötarkoitukset

Tähän mennessä olemme käsitelleet hakurobottia ainoastaan osana hakukonetta, mutta sillä on monia muitakin käyttötarkoituksia. Liikemailmassa hakurobottien keräämää dataa käytetään esimerkiksi hintojen sekä markkinatrendien mittaamiseen. Esimerkiksi verkkokauppaan keskittyvät web-sivut saattavat käyttää hakurobotteja keräämään tietoa kilpailijoiden tuotteiden hinnoista, ja sen jälkeen hinnoitella dynaamisesti omat tuotteensa hieman halvemmiksi.

(Khder 2021.) Hakuroboteilla tehtävä tiedonharavointi on suuressa suosiossa myös tutkijoiden keskuudessa, sillä se mahdollistaa suuren datamäärän keräämisen nopeasti ja halvasti. Esimerkiksi Liu ja Zhao (2017) hyödynsivät hakurobottia kerätäkseen Weibo-nimisestä sosiaalisen median palvelusta dataa kiinalaisten mielipiteistä liittyen ilmastonmuutokseen. Vihamieliset toimijat voivat käyttää hakurobotteja keräämään verkosta ihmisten henkilökohtaista informaatiota, jota voidaan myöhemmin käyttää esimerkiksi häiritsevien sähköpostilistojen muodostamiseen tai jopa identiteettivarkauksiin (Giles, Sun ja Councill 2010).

## 2.3 Robots Exclusion -protokolla

Robots Exclusion -protokolla on alalle vakiintunut epävirallinen standardi, jota eettiset hakurobotteja käyttävät toimijat noudattavat. Protokollan avulla web-sivun omistaja voi esimerkiksi ohjeistaa hakurobotteja tallentamasta sivustoa, tai tiettyjä sen osia. Protokolla on määritelty vuonna 1994 (Koster 1994) ja kuvattu tarkemmin RFC-dokumenttina vuonna 1996 (Koster 1996). Määritelmää on laajennettu, ja samalla ehdotettu viralliseksi standardiksi vuonna 2022 Googlen kanssa yhteistyössä julkaistussa RFC-dokumentissa (Koster ym. 2022). Käytännössä protokolla toteutetaan Robots.txt-tiedoston avulla.

# 3 Robots.txt tiedosto

Robots.txt-tiedosto tallennetaan web-sivujen juurihakemistoon. Hakurobotteja koskevat säännöt määritellään formaatiltaan seuraavanlaisina rivinvaihdoksin eroteltuina tekstiriveinä: "Kentän nimi : arvo". (Koster 1994.) Robots.txt tiedostossa voidaan määritellä seuraavia kenttien arvoja: User-agent, Disallow, Allow, Sitemap ja Crawl-delay.

## 3.1 Kenttien esittely ja syntaksi

Robots.txt-tiedosto koostetaan joko yhdestä sääntökokonaisuudesta, jota kaikki sivulla vierailevat hakurobotit noudattavat, tai sitten voidaan voidaan koostaa useampia eri sääntökokonaisuuksia säätelemään eri hakurobotteja. Robotin tulee ilmaista oma nimensä HTTP GET-pyynnön mukana tulevassa `user-agent`-otsikossa. Sääntöjen määrittely aloitetaan `user-agent`-kentästä, jossa määritellään minkä nimistä/nimisiä robotteja kentän alla olevat säännöt koskee. (Koster 1994.) Esimerkiksi rivi "user-agent: googlebot"tarkoittaa kyseisen rivin alla olevien sääntöjen koskevan Googlen yleistä hakurobottia (Google, n.d.). Mikäli robots.txt-tiedostossa on määritelty useampi kuin yksi kokoelma sääntöjä saman nimisellä `user-agent`-kentällä, tulee kyseisen robotin yhdistää kaikki kyseiset sääntökokoelmat yhdeksi ja noudattaa sitä. Mikäli halutaan kaikkien sivulla vierailevien robottejen noudattavan yhtä sääntörypästä, voidaan Robots.txt tiedosto koostaa vain yhdestä sääntöryppäästä jonka `user-agent` kentän arvoksi asetetaan asteriski "*". Asteriskilla varustettua `user-agent`-kenttää käytetään yleisesti myös määrittelemään kaikkien niiden hakurobottien sääntöjä, jotka eivät löytäneet omaa nimeään aiemmista `user-agent`-kentistä. Mikäli tietyn nimiselle hakurobotille ei ole määritelty omaa sääntöryhmää eikä tiedostossa ole asteriskilla varustettua `user-agent`-kenttää, tällöin kyseistä robottia ei rajoita mitkään säännöt.(Koster 1994.)

Kenttien allow ja disallow arvoilla arvolla säädellään mitä osia web-sivuston rakenteesta hakurobotti saa tai ei saa hakea (Koster ym. 2022). Kenttien arvoksi asetetaan hakemiston/sivun suhteellinen tiedostopolku verkkotunnukseen (engl. domain) nähden. Esimerkiksi rivi "Disallow: /salainenkansio"kieltää hakemasta kansion sisältämiä tiedostoja. `Allow-`

kentällä voidaan vastaavasti sallia tiettyjen polkujen hakemista. Mikäli jotain tiedostopolkua ei ole määritelty kielletyksi tai sallituksi, hakurobotti olettaa, että se on sallittua hakea ja tallentaa.

Robots.txt:n `allow`-kenttä sekä jokerimerkkien käyttö (muualla kuin `user-agent`-kentässä) ovat esimerkkejä siitä, kuinka protokollan viralliseen määritelmään ollaan myöhemmin lisätty suurimpien hakukoneiden harjoittamia vakiintuneita käytäntöjä. `Crawl-delay` ja `sitemap` -kentät ovat esimerkkejä kentistä jotka ovat virallisen määritelmän ulkopuolella, mutta silti käytössä ainakin joillain suurilla toimijoilla.

`Crawl-delay`-kentän avulla hakurobottia voidaan ohjeistaa odottamaan kentän arvon (positiivinen kokonaisluku) verran sekunteja jokaisen web-sivulle tehtävän pyynnön välissä. `Crawl-delay`-kentän käyttäminen voi olla tietyissä tilanteissa toivottavaa, sillä hakurobottien pyyntöjen hidastaminen vähentää web-sivua ylläpitävän serverin kuormaa. Isoista toimijoista ainakin Microsoftin Bingbot ilmoittaa dokumentaatiossaan kunnioittavan `crawl-delay`-kenttää. (Microsoft, n.d.). On olemassa myös muita strategioita varmistamaan ettei hakupyyntöjä tehdä liian nopeasti. Esimerkiksi Googlen hakurobotit käyttävät erilaisia algoritmeja määrittelemään sopivaa hakunopeutta sivun kantokykyyn nähden. Lisäksi Google tarjoaa mahdollisuuden säädellä verkkosivuston hakunopeutta heidän oman verkkosivun kautta. (Google, n.d.)

`Sitemap`-kentän tarkoituksena on auttaa hakurobottia suorittamaan web-sivuston haku järkevämmin (Schonfeld ja Shivakumar 2009). `Sitemap`-kenttään asetetaan tiedostopolku XML-tiedostoon, joka sisältää kuvauksen web-sivun rakenteesta sekä metatietoa jokaisesta sen sivusta. Web-sivun eri tiedostopolkujen lisäksi XML-tunnisteilla voidaan kertoa esim. kuinka usein sivua päivitetään, milloin se on viimeksi päivitetty sekä sivun priotiteettinumero. (Cyganiak ym. 2008.) `Sitemap`-kenttää käyttävät erityisesti sellaiset web-sivujen ylläpitäjät, jotka pyrkivät hakukoneoptimoimaan sivustoaan, sillä sitemap tiedoston avulla hakurobotille on mahdollista kertoa mitä sivuston osia hakurobotin erityisesti halutaan hakevan (Gregurec ja Grd 2012).

## 3.2 Robots Meta -tunniste ja muita menetelmiä

Robots Exclusion -protokollan lisäksi hakurobotteja voi ohjeistaa olla hakematta sivua (NOFOLLOW) tai olla hakematta sivustolta löytyviä linkkejä (NOINDEX) käyttämällä HTML merkintäkielen sekaan sijoitettavia meta tunnisteita. Meta-tunniste on syntaksiltaan seuraavanlainen: «META NAME="ROBOTS"CONTENT="NOINDEX, NOFOLLOW». (Yang ja Liao 2010.) Meta-tunnisteiden avulla myös sellainen web-sivuston julkaisija kenellä ei ole muokkausoikeuksia koko sivuston juurihakemistoon voi säädellä saako robotti hakea hänen julkaisemaansa sivua. Muita järeämpiä menetelmiä robottien estämiseksi on esimerkiksi IP-blokkaus (Chiapponi ym. 2022) ja CAPTCHA-testit (Zhang ym. 2022).

# 4 Haasteet

Web-sivustojen ylläpitäjien on syytä huomioida, että robots.txt- tiedoston lukeminen sekä kunnioittaminen on hakuroboteille täysin vapaaehtoista. Koster toteaakin alkuperäisessä vuoden 1994 määritelmässään, että "protokollan käyttöä ei valvo kukaan, eikä ole takuita siitä, että kaikki nykyiset sekä tulevat robotit kunnioittavat sitä". Siksi Robots.txt tiedostoa ei tule käyttää tapana hallinnoida sitä kenellä on, ja kenellä ei ole oikeutta päästä käsiksi nettisivun resursseihin (Koster).

Robots Exclusion -protokolla, kuten moni muukin edelleen käytössä olevista internet-protokollista on kehitetty vastaamaan aikansa tarpeisiin sekä haasteisiin. Nykyään hakurobotteja käyttää tutkijoiden ja internetasiantuntijoiden lisäksi monet kaupalliset toimijat sekä myös yksityishenkilöt. Internetissä on ladattavissa monia ilmaisia hakurobottiohjelmia, joita kuka tahansa pystyy lataamaan sekä käyttämään. Datan kasvanut arvo, hakukoneiden ja robottien yleistyminen sekä uudet innovaatiot ovat luoneet hakurobotteihin liittyviä ongelmia, joita 90-luvulla protokollien kehittäjät eivät osanneet vielä arvata. Thelwall ja Stuart (2006) esittävät 4 eri kategoriaa joihin hakurobottejen yhteiskunnalle tai yksityishenkilöille aiheuttamat ongelmat voidaan jakaa: palvelunesto, kustannukset, yksityisyys sekä tekijänoikeudet.

## 4.1 Palveleunesto

Mikäli verkkosivulle kohdistuu liikaa hakurobottien liikennettä, se voi hidastaa sivun käyttökokemusta muille käyttäjille (Chiapponi ym. 2022). Palvelunesto-ongelmat, jotka johtuvat web-palvelimien ylikuormituksesta olivat suurempi ongelma internetin alkuajoilla 90-luvulla kuin nykyään. Tämä johtuu tietokoneiden laskentatehon sekä internetin kaistan huomattavista parannuksista. Hakurobotit voivat kuitenkin edelleen kuormittaa etenkin pieniä web-servereitä sekä esimerkiksi kehittyvien maiden hitaampia tietoliikenneyhteyksiä. (Thelwall ja Stuart 2006). Koster (1993) kehoittaakin ”Guidelines for robot writers”-dokumentissa miettimään kahdesti, onko hakurobotin käyttäminen välttämätöntä päämäärän saavuttamiseksi. Myös web-sivujen robottiystävällisellä suunnittelulla, esimerkiksi välttämällä robottiansojen muodostamista (Heydon ja Najork 1999), voidaan vähentää robottien aiheuttamaa

turhaa kuormitusta. Kuormitusta voidaan vähentää myös käyttämällä aiemmin mainittua sitemap protokollaa sekä luvussa 3.1 esiteltyjä liiallisen kuormituksen estäviä algoritmeja.

## 4.2 Kustannukset

Web-servereiden palveluntarjoajat laskuttavat verkkosivujen ylläpitäjiä heidän alustansa käytöstä sivuston saaman liikenteen mukaan. Siksi hakurobottien turhasta liikenteestä voi aiheutua ylimääräisiä maksuja sivustojen ylläpitäjille. Toisaalta hakurobotit voivat myös tuoda uusia vierailijoita sivulle, jolloin robotin liikenteestä aiheutuvat maksut voidaan katsoa oikeutetuiksi. (Thelwall ja Stuart 2006.) Tiedonharavointia suorittavat hakurobotit voivat kuitenkin myös luoda tilanteen, jossa tiedonhaun kohteena olevan verkkosivun rahallinen arvo vähenee. Kerätty data voidaan julkaista uudelleen toisella sivulla, jolloin alkuperäisen datan julkaisijat eivät pääse hyötymään vierailijoista, jotka muuten synnyttäisi heille mainoksista saatavia tuloja. (Hirschey 2014.) Robottien torjumiseen keskittyvän yrityksen Datadomen mukaan tiedonharavoinnista johtuvien tulojen menetys saattaa olla jopa 10 prosenttia websivun kokonaisarvosta (Chiapponi ym. 2022).

## 4.3 Yksityisyys

Vaikka lähtökohtaisesti kaikki Internettiin sijoitettu tieto on julkista tietoa, hakurobotit voivat silti aiheuttaa ongelmia henkilöiden yksityisyydensuojalle. Yhdistelemällä dataa jota on kerätty eri Internet-sivustoilta, on mahdollista yhdistää data tiettyyn henkilöön (Mancosu ja Vegetti 2020). Hakurobottien myötä nettisivulle julkaistua henkilökohtaista tietoa saattaa päätyä kolmansille osapuolille, vaikka henkilökohtaisen tiedon julkaisija ei tätä olisi halunnut (Krotov ja Silva 2018). Myös hakurobottien avulla koostetut häiritsevät sähköpostilistat ovat yksi esimerkki sähköpostien haltijoiden yksityisyyden loukkaamisesta (Sun 2008).

## 4.4 Tekijänoikeudet

Osa verkkosivujen sisällöstä on suojeltu tekijänoikeuslailla. Siksi niiden kopioiminen hakurobotin avulla ilman sisällön omistajan lupaa saattaa loukata tekijänoikeuslakia (Thelwall

ja Stuart 2006). Lakiteknisistä syistä tiedonharavoijat joutuvat kuitenkin harvoin vastuuseen tekijänoikeusrikkomuksista. Esimerkiksi se, omistaako tiedon julkaissut nettisivu tekijänoikeudet itse tietoon on usein kyseenalaista. (Dreyer ja Stockton 2013.) Robots.txt tiedostoa on tulkittu myös oikeussaleissa, kun osapuolet ovat kiistelleet tekijänoikeusasioista. Robots.txt:n voidaan ajatella edustavan web-sivun julkaisijan toivomuksia siitä, saako sivun sisältämää dataa kopioida ja levittää. (Yang ja Liao 2010.)

# 5 Yhteenveto

Tässä tutkielmassa esiteltiin miten hakurobotti toimii, millaisiin käyttötarkoituksiin sitä käytetään sekä miten hakurobottien toimintaa voidaan säädellä Robots Exclusion -protokollan avulla. Esille tuotiin myös ongelmia, joita hakurobottien toiminta voi aiheuttaa, varsinkin jos robottien kehittäjät eivät kunnioita Robots.txt-tiedostossa määriteltyjä sääntöjä.

Tällä hetkellä epävirallisen standardin tulkintaerot aiheuttavat päänvaivaa sekä hakurobottien kehittäjille että verkkosivustojen ylläpitäjille. Siksi tällä hetkellä IETF:än (Internet Engineering Task Force) arvioitavana oleva ehdotus Robots Exclusion -protokollan muuttamisesta viralliseksi standardiksi olisi erittäin tervetullut muutos alalle.

Erityisesti tiedonharavointia suorittavien hakurobottien torjunnalle on kysyntää liikemaailmassa. Tästä syystä 2010-luvun loppupuolella on syntynyt ns. anti-bot-yrityksiä, jotka keskittyvät vähentämään robottien liikennettä asiakkaidensa verkkosivuilla. Mitä parempia tiedonharavoinnin torjuntatapoja kehitetään, sitä kekseliäämmin robotit tehdään kiertämään näitä esteitä. Hyvä esimerkki tästä on ns. CAPTCHA-farmit, tuore ilmiö jossa ihmisiä palkataan ratkaisemaan CAPTCHA- tehtäviä robottien puolesta. Näin verkkosivua voidaan huijata, että sivulla vierailija ei ole robotti. Tämä anti-bot teknologian sekä kekseliäiden hakurobottien suunnittelijoiden välinen kissa ja hiiri -leikki jatkunee myös tulevaisuudessa.

Vaikka tutkielmassa keskityttiin hakurobottien aiheuttamiin haasteisiin, haluaa kirjoittaja painottaa, että nykyisenkaltainen World Wide Web jossa kaikki tieto on heti saatavilla, ei olisi mahdollinen ilman hakurobottien tekemää työtä. Ennen automatisoituja hakurobotteja, 90-luvun alussa suosituimman hakukoneen Yahoon hakemistoja ylläpidettiin manuaalisesti käyttäjien toimesta. Tämä ei olisi enään mahdollista Internetin kasvaneen koon takia. Tästä syystä voidaan sanoa hakurobottien nettovaikutuksen olevan enemmän positiivinen kuin negatiivinen.

# Lähteet


Brin, Sergey, ja Lawrence Page. 1998. “The anatomy of a large-scale hypertextual web search engine”. *Computer networks and ISDN systems* 30 (1-7): 107–117.

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins ja Janet Wiener. 2000. “Graph structure in the web”. *Computer networks* 33 (1-6): 309–320.

Chiapponi, Elisa, Marc Dacier, Olivier Thonnard, Mohamed Fangar, Mattias Mattsson ja Vincent Rigal. 2022. “An industrial perspective on web scraping characteristics and open issues”. Teoksessa *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S),* 5–8. IEEE.

Cho, Junghoo, ja Hector Garcia-Molina. 2000. “Synchronizing a database to improve freshness”. *ACM sigmod record* 29 (2): 117–128.

Cothey, Viv. 2004. “Web-crawling reliability”. *Journal of the American Society for Information Science and Technology* 55 (14): 1228–1238.

Cyganiak, Richard, Holger Stenzhorn, Renaud Delbru, Stefan Decker ja Giovanni Tummarello. 2008. “Semantic sitemaps: Efficient and flexible access to datasets on the semantic web”. Teoksessa *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings 5,* 690–704. Springer.

Dreyer, A, ja Jamie Stockton. 2013. “Internet “data scraping”: A primer for counseling clients”. *New York Law Journal* 7:1–3.

Eichmann, David. 1994. “The RBSE spider-balancing effective search against web load”. Teoksessa *Proc. 1st WWW Conf.* Citeseer.

Giles, C Lee, Yang Sun ja Isaac G Councill. 2010. “Measuring the web crawler ethics”. Teoksessa *Proceedings of the 19th international conference on World wide web,* 1101–1102.

Google. n.d. “Change Googlebot crawl rate”. Viitattu 13. maaliskuuta 2023. https://support.google.com/webmasters/answer/48620?hl=en.

———. n.d. “How to write and submit a robots.txt file”. Viitattu 13. maaliskuuta 2023. https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt.

Gray, Matthew. n.d. “Credits and background”. Viitattu 13. maaliskuuta 2023. http://www.mit.edu.ezproxy.jyu.fi/people/mkgray/net/background.html.

Gregurec, Iva, ja Petra Grd. 2012. “Search Engine Optimization (SEO): Website analysis of selected faculties in Croatia”. Teoksessa *Proceedings of Central European Conference on Information and Intelligent Systems,* 211–218.

Heydon, Allan, ja Marc Najork. 1999. “Mercator: A scalable, extensible web crawler”. *World Wide Web* 2 (4): 219–229.

Hirschey, Jeffrey Kenneth. 2014. “Symbiotic relationships: Pragmatic acceptance of data scraping”. *Berkeley Tech. LJ* 29:897.

Kausar, Md Abu, VS Dhaka ja Sanjeev Kumar Singh. 2013. “Web crawler: a review”. *International Journal of Computer Applications* 63 (2): 31–36.

Khder, Moaiad Ahmad. 2021. “Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application.” *International Journal of Advances in Soft Computing & Its Applications* 13 (3).

Koster, M., G. Illyes, H. Zeller ja L. Sassman. 2022. *Robots Exclusion Protocol.* RFC9309. https://www.rfc-editor.org/rfc/rfc9309.txt.

Koster, Martijn. 1993. “Guidelines for robots writers” (tammikuu).

———. 1994. “A Standard for Robot Exclusion”. Viitattu 13. maaliskuuta 2023. http://www.robotstxt.org/orig.html.

———. 1996. *A Method for Web Robots Control.* https://www.robotstxt.org/norobots-rfc.txt.

———. *Surely listing sensitive files is asking for trouble?* https://www.robotstxt.org/faq/nosecurity.html.

Krotov, Vlad, ja Leiser Silva. 2018. “Legality and ethics of web scraping”. *Twenty- fourth Americas Conference on Information Systems, New Orleans.*

Kumar, Manish, Rajesh Bhatia ja Dhavleesh Rattan. 2017. “A survey of Web crawlers for information retrieval”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7 (6): e1218.

Liu, John, ja Bo Zhao. 2017. “Who speaks for climate change in China? Evidence from Weibo”. *Climatic Change* 140 (helmikuu). https://doi.org/10.1007/s10584-016-1883-y.

Mancosu, Moreno, ja Federico Vegetti. 2020. “What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data”. *Social Media+ Society* 6 (3): 2056305120940703.

McBryan, Oliver A. 1994. “GENVL and WWWW: Tools for taming the web”. Teoksessa *Proceedings of the first international world wide web conference,* nide 341. Citeseer.

Microsoft. n.d. “Crawl delay and the Bing crawler, MSNBot”. Viitattu 13. maaliskuuta 2023. https://blogs.bing.com/webmaster/2009/08/10/crawl-delay-and-the-bing-crawler-msnbot.

Pandey, Sandeep, ja Christopher Olston. 2005. “User-centric web crawling”. Teoksessa *Proceedings of the 14th international conference on World Wide Web,* 401–411.

Pinkerton, Brian. 1994. “Finding what people want: Experiences with the WebCrawler”. Teoksessa *Proc. of the 2nd Int. World Wide Web Conf., 1994.*

Schonfeld, Uri, ja Narayanan Shivakumar. 2009. “Sitemaps: above and beyond the crawl of duty”. Teoksessa *Proceedings of the 18th international conference on World wide web,* 991–1000.

Sun, Yang. 2008. “A comprehensive study of the regulation and behavior of web crawlers”. Tohtorinväitöskirja, The Pennsylvania State University.

Thelwall, Mike, ja David Stuart. 2006. “Web crawling ethics revisited: Cost, privacy, and denial of service”. *Journal of the American Society for Information Science and Technology* 57 (13): 1771–1779.

Udapure, Trupti V, Ravindra D Kale ja Rajesh C Dharmik. 2014. “Study of web crawler and its different types”. *IOSR Journal of Computer Engineering* 16 (1): 01–05.

W3C. n.d. “What is the difference between the Web and the Internet”. Viitattu 13. maaliskuuta 2023. https://www.w3.org/Help/#webinternet.

Yang, Chyan, ja Hsien-Jyh Liao. 2010. “Using the Robots. txt and Robots Meta tags to implement online copyright and a related amendment”. *Library hi tech* 28 (1): 94–106.

Yu, Linxuan, Yeli Li, Qingtao Zeng, Yanxiong Sun, Yuning Bian ja Wei He. 2020. “Summary of web crawler technology research”. Teoksessa *Journal of Physics: Conference Series,* 1449:012036. 1. IOP Publishing.

Zhang, Ning, Mohammadreza Ebrahimi, Weifeng Li ja Hsinchun Chen. 2022. “Counteracting dark Web text-based CAPTCHA with generative adversarial learning for proactive cyber threat intelligence”. *ACM Transactions on Management Information Systems (TMIS)* 13 (2): 1–21.