Lappeenrannan teknillinen yliopisto
School of Business and Management
Tietotekniikan koulutusohjelma

Kandidaatintyö

Karri Väänänen

VERKKOSIVUSTON KÄYTÖN MITTAAMINEN

Työn tarkastaja(t): TkT Uolevi Nikula

Työn ohjaaja(t): TkT Uolevi Nikula

Päiväys: 29.2.2016

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto
School of Business and Management
Tietotekniikan koulutusohjelma

Karri Väänänen

Verkkosivuston käytön mittaaminen

Kandidaatintyö, 2016
47 sivua, 11 kuvaa, 3 taulukkoa, 3 liitettä
Työn tarkastaja: TkT Uolevi Nikula
Hakusanat: kävijäseuranta, web-analytiikka, lokien analysointi, verkkosivuston tiedonlouhinta, Google Analytics, i+, iplus, Visit Lappeenranta
Keywords: user tracking, web analytics, log analysis, web mining, Google Analytics, i+, iplus, Visit Lappeenranta

Tässä työssä käsitellään kävijäseurannan menetelmiä ja toteutetaan niitä käytännössä. Web-analytiikkaohjelmistojen toimintaan tutustutaan, pääasiassa keskittyen Google Analyticsiin. Tavoitteena on selvittää Lappeenrannan matkailulaitepäätteiden käyttömääriä ja eriyttää niitä laitekohtaisesti. Web-analytiikasta tehdään kirjallisuuskatsaus ja kävijäseurantadataa analysoidaan sekä vertaillaan kahdesta eri verkkosivustosta. Lisäksi matkailulaitepäätteiden verkkosivuston lokeja tarkastellaan tiedonlouhinnan keinoin tarkoitusta varten kehitetyllä Python-sovelluksella.

Työn pohjalta voidaan todeta, ettei matkailulaitepäätteiden käyttömääriä voida nykyisen toteutuksen perusteella eriyttää laitekohtaisesti. Istuntojen määrää ja tapahtumia voidaan kuitenkin seurata. Matkailulaitepäätteiden kävijäseurannassa tunnistetaan useita ongelmia, kuten päätteiden automaattisen verkkosivunpäivityksen tuloksia vääristävä vaikutus, osittainen Google Analytics -integraatio ja tärkeimpänä päätteen yksilöivän tunnistetiedon puuttuminen. Työssä ehdotetaan ratkaisuja, joilla mahdollistetaan kävijäseurannan tehokas käyttö ja laitekohtainen seuranta. Saadut tulokset korostavat kävijäseurannan toteutuksen suunnitelmallisuuden tärkeyttä.

ABSTRACT

Lappeenranta University of Technology
School of Business and Management
Degree Program in Computer Science

Karri Väänänen

Measuring website usage

Bachelor's Thesis
2016
47 pages, 11 figures, 3 tables, 3 appendices
Examiner: D.Sc. Uolevi Nikula
Keywords: user tracking, web analytics, log analysis, web mining, Google Analytics, i+, iplus, Visit Lappeenranta

This thesis discusses and implements methods used in web analytics. The practicalities of web analytics software are made familiar with focus on Google Analytics. The goal is to report the usage of travel information terminal devices in Lappeenranta, and to differentiate the usage data between the devices. Literature study is done about web analytics and web analytics data from two different websites is analyzed. Moreover, the website log data of the devices are examined using data mining with a Python application developed for the purpose.

Based on the work it can be concluded that differentiating the usage data cannot be done from the current implementation. However, the number of sessions and events can be tracked. The implementation of web analytics in these information terminal devices have several problems, such as, skews in results caused by automatic page refresh, partial Google Analytics integration, and most importantly the missing unique identifiers. The thesis suggests solutions for enabling the efficient use of web analytics, and tracking the devices independently. The results emphasize the importance of a systematical web analytics implementation.

ALKUSANAT

Tämä kandidaatin tutkinnon opinnäytetyö on tehty Lappeenrannan teknillisen yliopiston tietotekniikan koulutusohjelmaan. Työn aloittamisen ja valmistumisen välille mahtui vaihto-opiskelujakso ja kesätöiden jatkaminen pitkälle syksyyn jokseenkin opiskelun ja tämän opinnäytetyön valmistumisen kustannuksella. Aikataulullisista haasteista huolimatta työ muotoutui harkitun kaltaiseksi. Kiitän työn ohjaajaa tuesta ja lähimpiä sukulaisiani mielenkiinnon osoittamisesta työtäni kohtaan.

SISÄLLYSLUETTELO

SYMBOLI- JA LYHENNELUETTELO

AJAX	Asyncronous JavaScript And XML
CBT	Cross-Border Travel
DNT	Do Not Track
DOM	Document Object Model
GA	Google Analytics
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTP BA	HTTP Basic Authentication
IP	Internet Protocol
JS	Javascript
KPI	Key Performance Indicator
PHP	PHP: Hypertext Preprocessor
UAS	User Agent String
URL	Uniform Resource Locator
UTM	Urchin Tracking Module

1 JOHDANTO

1.1 Tausta

Web-analytiikka eli kävijäseuranta on verkkosivulla kävijöiden toiminnan seuraamista ja tapahtumien keräämistä tietokantaan. Siihen kuuluu lisäksi kerätyn datan analysointi ja raportointi tuottaen informaatiota ja uutta tietoa verkkosivuston käytöstä. Kävijäseurannasta saadaan runsaasti erityyppistä dataa, joka ei välttämättä ole itsessään varsin hyödyllistä. Web-analytiikka on tämän semanttisesti merkityksettömän datan erittelyä, eri suhteiden ymmärtämistä ja analysointia, jotta aineistoista olisi mahdollisimman paljon hyötyä sitä tuottavan tahon toiminnalle. [1]

Kävijäseurannan tuottamaa tietoa voidaan käyttää moniin eri tarkoituksiin. Kävijäseurannan avulla voidaan esimerkiksi ymmärtää kuinka hyvin verkkosivusto palvelee sen käyttäjiä ja kuinka käyttökokemusta voidaan parantaa. Kävijäseuranta on myös tärkeä työkalu verkkokaupan myynnin ja asiakkaiden hankkimisen lisäämiseen tai minkä tahansa muun strategisen päämäärän saavuttamiseen. Markkinointia verkkomedioissa voidaan seurata kävijäseurannan avulla, sillä usein on tärkeää tietää mistä kävijä saapui verkkosivulle ja johtiko markkinointikampanja lopulta myynnin nousuun. Näiden käyttötapausten lisäksi web-analytiikka tarjoaa tietoa verkkosivuston suorituskyvystä ja mahdollisista ongelmista, kuten kuolleista hyperlinkeistä. [2]

Kävijäseurannan toteuttamiseen on useita eri vaihtoehtoja ja lähestymistapoja. Verkkosivustojen käyttöä voidaankin nykyään seurata monella tapaa. Kävijäseurantaan on olemassa useita valmiita palveluita, kuten Google Analytics, Yandex Metrica, StatCounter, TNS Metrix. Myös palvelimelle asennettavia ohjelmistoja on saatavilla kuten avoimen lähdekoodin Piwik.

Kävijäseurannassa ilmenee myös monia ongelmia, kuten seurantadatan tarkkuus tai huolet käyttäjien yksityisyydestä [2]. Käyttäjien seuraaminen heidän tietämättään on oleellinen huomioonotettava asia verkkosivuja kehitettäessä. Sähköisen viestinnän tietosuojalaki

velvoittaakin palveluntarjoajat antamaan käyttäjälle mahdollisuuden kieltää heidän tietojensa seuraamiseen tarkoitettujen evästeiden tallentaminen. [3]

1.2 Tavoitteet ja rajaukset

Työssä tehdään ennen kaikkea kirjallisuuskatsaus verkkosivuston käytön mittaamista käsitteleviin tutkimuksiin sekä julkaisuihin. Katsotaan saatavilla olevan aineiston määrä ja millaista aineisto on. Tärkeää on siis tutustua tehtyyn tutkimukseen aiheesta ja miten verkkosivun käyttöä on seurattu aiemmin.

Tarkoituksena on tutustua verkkosivuston käytön mittaamisen historiaan ja selvittää millaisilla työkaluilla verkkosivustojen käyttöä seurataan. Miten nämä työkalut toimivat käytännössä ja millaista tietoa ne tuottavat sovelluskehittäjien sekä markkinoinnin avuksi. Eri työkaluja pyritään vertailemaan ja selvittämään mihin niiden tuottamaa tietoa voidaan käyttää. Työssä luodaan katsaus molempiin palvelimille asennettaviin ratkaisuihin sekä verkkosivun merkitsemiseen perustuviin kolmannen osapuolen tarjoamiin palveluihin kuten Google Analytics. Selvitetään tekniikat työkalujen takana ja miten käyttäjiä identifioidaan.

Työssä analysoidaan kahdesta verkkosivustosta saatua dataa. Google Analytics -metriikkaa saadaan analysoitavaksi Cross-Border Travel -sivustolta, joka on Suomen, Venäjän ja Euroopan Unionin yhdessä rahoittama palvelu rajamatkailun edistämiseksi. Tämän lisäksi myös Lappeenrannan matkailun ylläpitämiltä i+-matkailuneuvontalaitteilta (i+-päätteet) on saatavissa seurantatietoja Google Analytics palvelussa, sekä Visit Lappeenranta -verkkosivuston palvelimen lokitiedoista. Tavoitteena näiden päätteiden osalta on selvittää päätteiden käyttömääriä yleisesti ja päätekohtaisesti.

1.3 Työn rakenne

Tämä työ käsittää kirjallisuuskatsauksen web-analytiikkaan. Seuraavissa luvuissa käydään läpi kirjallisuudesta löydettyjä eri menetelmiä kävijäseurannassa, sekä selvitetään osin näiden menetelmien käytettävyyttä käytännössä tämän työn osana sekä yleisesti yrityksen verkkosivustoa mittaavina työkaluina.

Taustatiedot luvussa kerrotaan lisää haastattelujen perusteella saaduista tiedoista tässä työssä käsiteltävistä järjestelmistä, tavoitteista kävijäseurannalle ja seurannan kohteista.

Työn toinen osa keskittyy kerättyjen seurantatietojen analysointiin erikseen crossbordertravel.fi ja i+-matkailuneuvontalaitteiden näkökulmasta. i+-päätteiden kohdalla analysoidaan sivujen merkitsemiseen perustuvin menetelmin saatua dataa, sekä sen lisäksi käydään läpi www-palvelimen lokien tuottamaa dataa. Kävijäseurantatyökaluihin, kuten Google Analytics, tutustutaan käytännössä. Lisäksi niistä saatavaa dataa käydään läpi ja analysoidaan.

1.4 Kirjallisuuskatsaus

Työn kirjallisuuskatsaus pohjautuu yli 10 tieteelliseen julkaisuun, muutamiin kävijäseurantaan liittyviin ohjelmistokohtaisiin ohjeisiin sekä muihin aihetta käsitteleviin verkkoartikkeleihin. Julkaisuissa käsitellään muun muassa verkkosivun tehokkuuden mittaamista [4], lokien analysointia [5], reaaliaikaisen ja ennakoivan analytiikan käyttämistä dynaamisen sisällön näyttämiseen [6], Google Analyticsin käyttöä verkkosivun käytettävyyden arviointiin [1], kävijäseurantatyökalun valintaa [7], hiiren liikkeiden seuraamista [8] ja kävijäseurannan soveltamista verkkoportaaliin [9]. Lisäksi yhdessä näistä julkaisuista käydään läpi case-tutkimuksen avulla käyttäjäseurannan soveltuvuutta akateemiseen kirjastoympäristöön ja mitä se vaatii käytännössä [10].

Julkaisuissa käsitellään myös miten datasta saatua tietoa voidaan käyttää markkinoinnin apuna ja käyttäjien tai asiakkaiden tunnistamisessa. Tällaisten papereiden aiheina olivat suuren tuoton antavien asiakkaiden tunnistaminen ja käyttäjien dynamiikan tutkimus kävijäseurannan avulla [11] [12]. Muita mielenkiintoisia aiheita löytyi myös, kuten kävijäseuranta työkaluna strategisessa kommunikoinnissa [13], käyttäjän tunnistus digitaalisen sormenjäljen avulla [14] ja käytönseuranta verkkoyhteisöissä [15].

Varsinaisia kirjoja opetustarkoitukseen tai muunlaista kirjallisuutta löytyi vähän. Muutamia oppaita työkalujen käyttöön oli saatavilla sekä useita sosiaalista mediaa käsitteleviä teoksia. Nämä olivat kuitenkin parhaimmillaankin vain työn aihetta sivuavia.

2 TAUSTATIEDOT

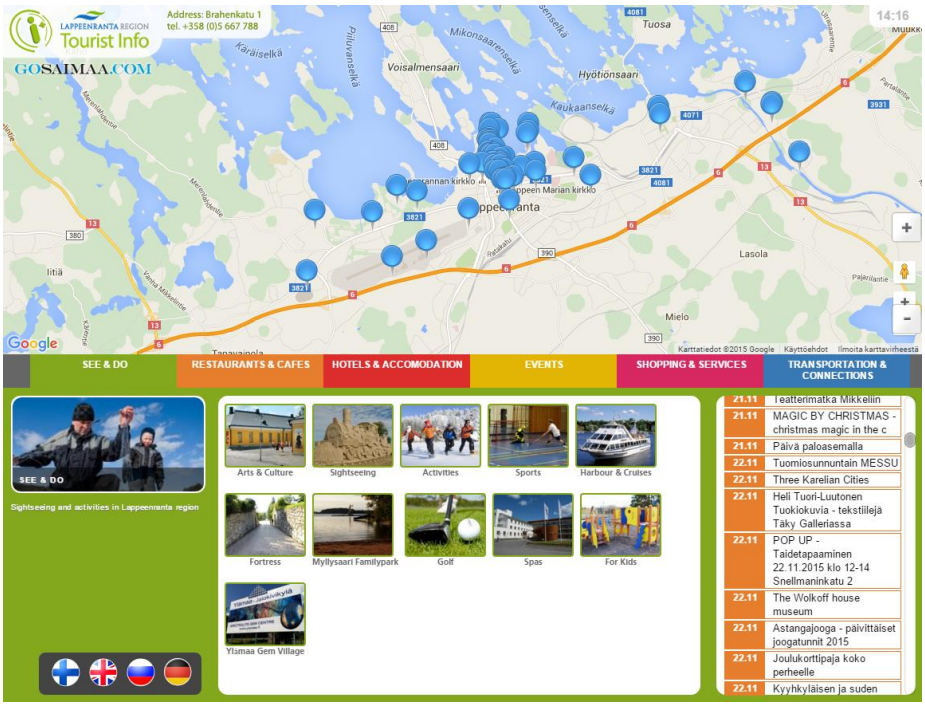
2.1 Lappeenrannan matkailun i+-matkailuneuvontalaitteet

Lappeenrannan matkailu (www.visitlappeenranta.fi) on Lappeenrannan kaupungin organisaatio ja verkkosivusto alueen matkailun kehittämiseksi. Lappeenrannan matkailu on hankkinut i+-järjestelmän, joka koostuu 18 eri päätteestä [16]. Päätteet ovat kosketusnäytöllä varustettuja yksinkertaistettuja Windows XP -koneita, joiden ainoa tarkoitus on näyttää i+-verkkosivustoa kioskitilassa eli näkymässä, jossa kaikki käyttöjärjestelmään liittyvät käyttöliittymän elementit on piilotettu käyttäjältä.

Kuva 1: Lappeenrannan matkailun i+-pääte

Järjestelmän on toimittanut Videra Oy yhdessä MediaNyt-mainostoimiston kanssa ja se on rakennettu Grassfish-ratkaisun päälle. Järjestelmää ylläpitää edellä mainittu Videra Oy, mutta Lappeenrannan matkailulla on mahdollisuus lisätä ja muokata sisältöä. Verkkosivuston alustana toimii InfoPro+-järjestelmä [17]. Järjestelmä päivittää päätteen näkymän automaattisesti ajoittain, mikä vaikeuttaa kävijäseurannan toteuttamista tässä ympäristössä.

i+-verkkosivustoa seurataan Google Analyticsin avulla. Verkkosivusto on hankittu alihankintana ja toteutettu CodeIgniter-sovelluskehyksen päälle PHP-kielellä (PHP: Hypertext Preprocessor) ja MySQL-tietokannan avulla. Tietoliikenne päätteiden ja palvelimen välillä on toteutettu pääasiassa AJAX-kutsuilla (Asynchronous JavaScript And XML) lukuun ottamatta päätteen selaimen tekemiä ajoittaisia sivun automaattisia uudelleenlatauksia. Tästä johtuen Google Analytics -toiminnallisuus on toteutettu käyttäen tapahtumaseurantaa. Sivun tapahtumien tallennus, kuten käyttäjän navigointi sivulla, vaatii tässä tapauksessa Analytics-sovellusrajapinnan syvällisempää integrointia verkkosivustoon. Käytännössä jokainen tapahtuma pitää erikseen rekisteröidä käyttäen kyseistä sovellusrajapintaa. [18]



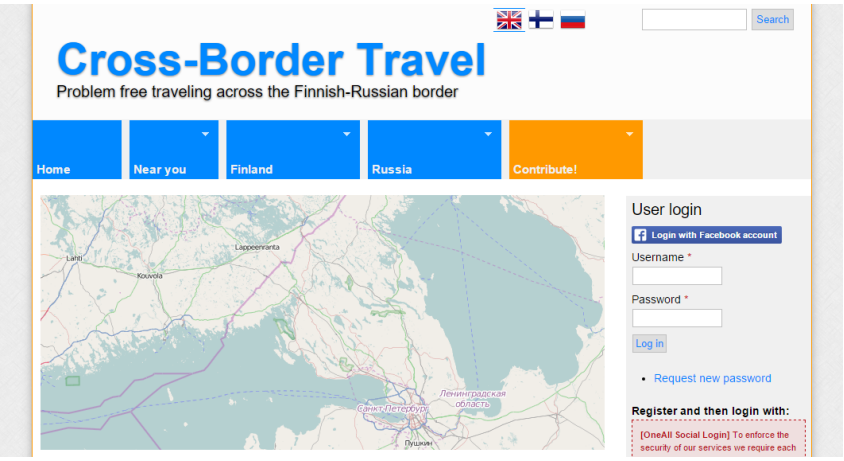
Kuva 2: i+-päätteen käyttöliittymä [19]

Nykyisen toteutuksen ongelmana on, että kävijämääriä ei seurata päätekohtaisesti, eikä Google Analyticsin tarjoamaa potentiaalia ei ole hyödynnetty. Google Analytics raportoima kävijämäärä on suurimmaksi osaksi vakio, noin 18–20 kävijää eli juurikin päätteiden lukumäärä. Vielä suurempana ongelmana on, ettei saatuja tapahtumia eritellä laitteiden

kesken, vaan kaikkien päätteiden tapahtumat lasketaan samaan kokonaisuuteen. Näin ei tiedetä onko jokin pääte, jolla on hyvin vähän käyttöä tai onko päätteitä, jotka ovat erityisen paljon käytettyä.

2.2 Cross-Border Travel -verkkosivusto

Cross-Border Travel on verkkosivusto osoitteessa www.crossbordertravel.eu. Sivusto on Suomen, Venäjän ja Euroopan Unionin yhdessä rahoittama palvelu rajamatkailun edistämiseksi. Tarkastelujakson aikana seurantatietoja oli kerätty Google Analyticsin avulla. Kyseessä on suhteellisen pieni verkkosivusto, joka on edelleen kehitteillä ja kuvassa 3 näkyvä versio otettiin pois käytöstä vuoden 2015 aikana. Cross-Border Travel tarjoaa vertailupohjan Lappeenrannan matkailun verkkosivustolle. Molemmat käyttävät samaa Google Analytics -kävijäseurantaa, minkä johdosta vertailu saatujen tulosten välillä on helppoa.



Kuva 3: CBT-verkkosivusto [20]

3 KÄVIJÄSEURANNAN MENETELMÄT

3.1 Web-analytiikkaohjelmistot

Yleisimmin web-analyytiikkaohjelmistot perustuvat sivujen merkitsemiseen (page tagging). Sivuille lisätään usein lyhyt JavaScript-koodi, joka kerää tiedot sivusta sekä käyttäjästä ja lähettää nämä tiedot verkon yli palvelimelle. [21] Kolmannen osapuolen tarjoamien seurantapalvelujen, kuten Google Analytics ja Yandex Metrica, lisäksi suoraan palvelimelle asennettavat työkalut pystyvät analysoimaan suoraan myös palvelinohjelmien lokeja. Piwik on esimerkiksi eräs suoraan palvelimelle asennettava ohjelmisto. Tällöin ei olla riippuvaisia ulkopuolisesta palvelusta ja kaikki seurantatieto on vain omassa käytössä. Tällaisen ratkaisun valinnalla voi olla merkitystä kun tietoturva on tärkeä valintaperuste. Piwik on ensisijaisesti sivujen merkitsemiseen perustuva ohjelmisto, mutta se tarjoaa myös vaihtoehtoisen ratkaisun lokien analysointiin. [22]

Kävijäseurantaohjelmistot tuottavat monenlaista tietoa, joka on usein hyödyllistä eri tarkoituksiin. Verkkosivun kävijäseuranta antaa mahdollisuudet verkkosivun kehittämiseen ja sovelluskehittäjille suunnatun tiedon tarjoamiseen, mutta myös toisaalta mainonnan, myynnin ja yleisesti markkinoinnin apuvälineeksi. [2]

3.2 Kävijäseurannan menetelmät

Web-analytiikassa seurantatietojen keräysmenetelmät voidaan jakaa neljään eri tyyppiin. Sivun merkitseminen (page tagging) on yleisin, mutta myös lokien analysointi, upotetut elementit (”web bug” tai ”web beacon”), sekä verkon tarkkailu ovat käytettäviä menetelmiä. [7] [1]

Sivun merkitseminen on menetelmä, jota useimmat nykyiset web-analytiikkapalvelut käyttävät. Siinä jokaiselle seurattavalle sivulle asetetaan lyhyt JavaScript-koodi, joka kerää tietoja käyttäjästä ja lähettää ne palveluntarjoajan palvelimelle. JavaScript-toteutuksella pystytään käsittelemään selaimen tarjoamaa DOM-puuta (Document Object Model) ja voidaan siten kerätä tietoja käyttäjän selaimesta, käyttöjärjestelmästä, kielestä,

selainlaajennuksista, aikavyöhykkeestä, näytön resoluutiosta ja värisyvyydestä, käyttöjärjestelmän fonteista ja ovatko evästeet sallittuja. [23] Lisäksi saadaan selville selaimen Flash tai Java -tuki. [21] [14]

Verkkosivupalvelimen lokien analysointi on puolestaan yksi ensimmäisiä web-analytiikan ilmenemismuotoja [24]. Ongelmana on kuitenkin kerätyn tiedon vähyys verrattuna JavaScriptillä saataviin tietoihin. Myös käyttäjien liikkeiden seuraaminen sivustolla on hankalaa ja hakukoneiden aiheuttama liikenne helposti vääristää tuloksia. Lisäksi dynaamiset, paljon JavaScript tai Flash -toteutusta sisältävät sivustot ovat ongelmallisia seurattavia. [5] Lokitiedot sisältävät usein käyttäjän IP-osoitteen (Internet Protocol), aikaleiman, pyynnön tiedot ja vastauksen tilakoodin lisäksi UAS-otsikkotiedon (User Agent String). UAS on asiakkaan selaimen lähettämä merkkijono, joka kertoo tärkeimmät tiedot järjestelmästä ja itse selaimesta [14].

Web bugit ovat kolmas seurantatapa. Ne ovat HTML-tiedostoon (Hypertext Markup Language) upotettuja elementtejä, jotka haetaan ulkoisesta lähteestä. Tyypillisesti web bugi on toteutettu yhden pikselin kokoisella läpinäkyvällä GIF-kuvalla (Graphics Interchange Format) img-elementissä, mutta se voidaan toteuttaa myös muilla HTML-elementeillä, kuten img, embed, iframe tai object. Web bugin toiminta perustuu siihen kun käyttäjä avaa sivun selaimellaan tai lukee HTML-muotoisen sähköpostin. Sivu ladataan ja web bug elementti haetaan verkon yli kolmannen osapuolen palvelimelta. Web bug -elementin URL:ään (Uniform Resource Locator) voidaan lisätä tietoa sen query-osaan [25] samaan tapaan kuin Google Analyticsin muokatuiden kampanjoiden linkkeihin lisätään UTM-parametreja[1] (Urchin Tracking Module) [26]. Usein web bugeja käytetään yhdessä evästeiden kanssa, jolloin voidaan yksilöidä käyttäjä ja seurata hänen etenemistään sivulla.

[1] UTM-parametrit juontavat nimensä Urchin-ohjelmistosta, jonka Google osti Urchin Software Corporation -yritykseltä vuonna 2005. Urchin-ohjelmistosta tuli siten myöhemmin perusta Google Analyticsille. [38]

Myös verkkopakettien tarkkailu voidaan katsoa olevan osa verkkosivuston käytönseurantaa. Käytännössä seurantamenetelmät kuitenkin jaetaan sivujen merkitsemiseen ja lokien analysointiin perustuviin. [7] Lokien analysoinnilla saatavan tiedon vähyys ja epätarkkuus johti kuitenkin sivujen merkitsemiseen perustuvien menetelmien yleistymiseen [1].

3.3 Google Analytics -seurantatiedon lähteet

Google Analytics -palvelu kerää kaikki seurantatiedot seuraavista lähteistä: HTTP-pyynnön (Hypertext Transfer Protocol) otsikkotiedot (Liite 1), selaimen tarjoama DOM-puu ja ensimmäisen osapuolen evästeet. [21] Esimerkkitarkasteluun otetusta Google Analyticsiä käyttävästä verkkosivustosta tallennettiin HTTP-pyynnöt käyttämällä Firefox 25.0 selainta, johon oli asennettu Firebug 1.12.4 -selainlaajennus.

Web bugin hakemiseen käytetyn HTTP-pyynnön muoto riippuu käytettävästä selaimesta ja lähetettäviin pyyntöihin voidaan tehdä muutoksia esimerkiksi käyttämällä selainlaajennuksia. Tyypillinen HTTP GET -pyyntö on esitetty liitteessä 1. Pyynnön mukana menee tieto halutusta resurssista, käytetystä selaimesta, käyttöjärjestelmästä, kielestä ja Referer-kentässä miltä sivulta käyttäjä saapui. Luonnollisesti saadaan selville myös käyttäjän IP-osoite verkkotason paketista sekä aikaleima. Analytics kerää tämän tiedon web bugin avulla, joka on osa useita menetelmiä, joita Analytics käyttää [21].

JavaScriptin avulla päästään käsittelemään selaimen tarjoamaa lisätietoa järjestelmästä ja DOM-puuta. Käyttäjän selaimessa Google Analytics toiminnallisuus on toteutettu ga.js tai analytics.js Javascript-tiedostossa, joka ladataan verkkosivun yhteydessä käyttäjän koneelle. Liitteessä 2 on esitetty HTTP-pyyntö, jolla ga.js lähettää seurantatiedot Googlen palvelimille. Tässä JavaScriptin avulla saatu tieto on tallennettu GIF-parametreina pyynnön query-osaan [21] [25]. GIF-parametreiksi kutsutaan yksinkertaisesti ”__utm.gif” -sivun hyväksymää joukkoa parametreja.

Kolmas tärkeä elementti Google Analytics -toteutuksessa on asiakkaan koneelle tallennettavat ensimmäisen osapuolen evästeet. Evästeiden perimmäisenä tarkoituksena on tunnistaa palaavat käyttäjät uusista, sekä tarjota mahdollisuus istuntojen seuraamiselle.

Uusimmassa analytics.js-koodissa tallennetaan kaksi evästettä, käyttäjän tunnistamiseen ja toinen pyyntömäärien hallitsemiseen. Käyttäjän tunnistamiseen tarkoitettu eväste poistuu käyttäjän koneelta automaattisesti vasta kahden vuoden kuluttua sen viimeisimmästä aktivoinnista. [27]

Google Analytics tarjoaa myös ”muokatut kampanjat” -palvelun. Siinä kampanjoita ja liikenteen lähteitä seurataan UTM-parametrien avulla. Google Analytics mahdollistaa liikenteen lähteen seurannan, esimerkiksi lisäämällä uutiskirjeessä esiintyviin linkkeihin tieto lähteestä ja mihin kampanjaan se liittyy. Esimerkiksi mainossähköposteihin oleviin linkkeihin voidaan lisätä UTM-parametreja tuottamaan lisätietoa liikenteen lähteestä ja kampanjasta, johon liikenne liitetään. [26] Toisin kuin GIF-parametrit nämä liitetään suoraan oman verkkosivun linkkeihin, josta Analytics-koodi välittää ne edelleen Googlen palvelimille.

3.4 Selainriippumaton kävijäseuranta ja virtuaalinen sormenjälki

Internetiin kytkettyjen laitteiden määrän kasvaessa ja yhä useampien käyttäjien käyttäessä useampaa kuin yhtä selainta, on selainriippumattomille seurantamenetelmille kysyntää. Evästeisiin perustuvat käyttäjän yksilöivät menetelmät eivät toimi, kun käytetään useampaa kuin yhtä selainta. Lisäksi evästeet voidaan poistaa tai estää helposti käyttäjän toimesta.

Käyttämällä hyväksi paikannettua IP-osoitetta, sovellustason informaatiota, kuten HTTP-otsikkotietoja, sekä Javascriptillä saatavaa tietoa järjestelmästä ja selaimesta, voidaan luoda yksilöllinen sormenjälki jokaiselle käyttäjälle. Saatu sormenjälki ei välttämättä riipu käytetystä selaimesta tai tarkasta IP-osoitteesta, joten selaimen vaihtaminen tai dynaamisen IP-osoitteen vaihtuminen ei vaikuta tunnisteeseen. [14]

Yksilöllisen tunnisteen eli virtuaalisen sormenjäljen mahdollistavat korkean entropian omaamat tietolähteet, kuten tarkat sovellusversiot ja pääsy käyttöjärjestelmän fontteihin tai selaimen laajennuksiin. Esimerkiksi Windows-järjestelmässä erilaisten fonttien määrä on hyvin korkea ja samoin niiden mahdolliset eri kombinaatiot. Näin edes välityspalvelimen käyttäminen ei takaa anonymiteettiä, sillä pelkän IP-osoitteen muuttuminen voidaan

huomata ja olettaa käyttäjän olevan sama. Myöskään evästeiden estäminen ei auta tällaisen passiivisen seurannan tapauksessa. Ainoastaan tarkoituksellinen selaimen lähettämien tietojen jatkuva muuttaminen tekee seuraamisen mahdottomaksi. Myös JavaScript-ominaisuuden poistaminen käytöstä heikentää tunnistetta sillä silloin menetetään suuri osa saatavasta tiedosta, jolla yksilöinti tapahtuu. [14]

3.5 Suorituskykymittarit kävijäseurannassa

Kävijäseurannasta saatua dataa analysoidaan monin eri tavoin ja siitä saatua tilastollista informaatiota kuvataan useilla eri suhdeluvuilla ja suorituskykymittareilla. Kuten esimerkiksi taloudellisia indikaattoreita myös kävijäseurannan keskeisiä mittareita kutsutaan usein KPI-mittaristoksi (Key Performance Indicator) [28] [10]. Näiden indikaattoreiden oikein tulkitseminen vaatii, että tiedetään mitä ne kuvaavat ja millä perusteella ne on laskettu.

Pelkästään hyvin yksinkertaisten mittareiden kuten sivunkatselujen määrän käyttäminen voi johtaa hyvin harhaanjohtaviin tuloksiin. Varsinkin jos turvaudutaan ainoastaan lokien analysointiin, ne eivät ota kantaa siihen oliko kyseessä oikeasti sivusta kiinnostunut ihminen vai vain hakurobotti. Tämä voi vääristää tuloksia merkittävästi. [11] [15]

Fagan, J.C., käy läpi artikkelissaan [10] kävijäseurannan metriikkaa. Artikkeli tunnistaa 24 eri suorituskykymittaria, joista useimmat ovat toteutettu muun muassa Google Analytics -palvelussa. Artikkeli huomauttaa myös, että käytettävät mittarit tulee valita käyttötapauksen mukaan. Käytettäviä mittareita ei tulisi myöskään valita kovin montaa, sillä organisaatiossa se vaikeuttaa hahmottamista sen suhteen mitä oikeastaan mitataan ja mikä organisaatiolle on oikeasti tärkeää.

3.6 Kävijäseurantatyökalun valintakriteerit

Oikean seurantatyökalun valinnassa käyttäjän erityispiirteet on otettava huomioon. Jälleenmyyjän ja tuotteita tuottavan yrityksen päämäärät ovat usein erilaiset verkkosivuston näkökannalta. Verkkokauppa pyrkii kasvattamaan myyntiä verkkosivustonsa kautta, kun

taas tuottava yritys voi pyrkiä ensisijaisesti informoimaan tuotteistaan, kertomaan yhteistietojaan tai asiakastuestaan. Seurantapalvelun valinnassa KPI:n kustomoitavuus ja esilletuonti esimerkiksi kojelauta-tyylisellä (engl. dashboard) ratkaisulla voi olla tärkeää. Toisaalta tutkimuksessa tuotiin esille, ettei informaation esitystavalla ole niinkään väliä vastaajien perusteella, mutta ulkonäölliset tekijät silti vaikuttivat työkalun valintaan. [7]

Tärkeimmät kolme valintaan vaikuttavaa tekijää ovat käytettävä seurantamenetelmä (kuten sivun merkitseminen tai lokien analysoiminen), asennustapa (palvelimelle asennettava tai ulkopuolinen palvelu) ja palvelun hinta. [7]

3.7 Seurannan ongelmat ja käyttäjän yksityisyys

Kävijäseuranta voi nostaa esille yllättäviäkin käyttäjien yksityisyyteen liittyviä ongelmia. Verkossa asioivien käyttäjien yksityisyyttä onkin pyritty parantamaan lainsäädännön avulla Suomessa ja EU:n alueella. Suomessa velvoitetaan verkkosivuja antamaan käyttäjille mahdollisuuksia kieltää heidän tietojensa seuraaminen [3]. Myös Euroopan Unionin tasolla verkkosivuja velvoitetaan ilmoittamaan käyttäjälle tapahtuvasta seurannasta [29].

Palveluntarjoajat ovat kehittäneet valinnaisia menetelmiä yksityisyyden parantamiseksi, kuten DNT (Do Not Track) HTTP-otsikkotiedon tuki [30] tai Google Analyticsin tarjoama IP-osoitteen anonymiteetti [31] [32]. Ongelmana on monesti tällaisten menetelmien vapaavalinnaisuus, niiden aktivoimisen vaikeus tai paikallisuus. Internet on luonteeltaan globaali, kun taas lait pätevät usein vain kovin paikallisesti.

Verkkosivuston seurantaan tarkoitetuilla palveluilla voi olla yksityisyyden kannalta varsin arveluttaviakin ominaisuuksia. Yandex Metrican tarjoama Webvisor kykenee tallentamaan jokaisen hiiren ja näppäimistön painalluksen ajan funktiona. Täten voidaan nauhoittaa käyttäjän toimet ja katsoa ne myöhemmin uudelleen. Verkkosivujen luonteen vuoksi on lähes sääntö, ettei käyttäjä huomaa tällaisen seurannan tapahtumista, jollei siitä erikseen ilmoiteta kyseisellä verkkosivulla. Webvisorin seurantamenetelmät voivat tallentaa myös käyttäjän syötteen mitä käyttäjä ei ollut alun perin tarkoittanut lähetettäväksi. Esimerkiksi, jos käyttäjä alkaa syöttämään virheellisesti salasanaa tunnuskenttään salasana lähetetään jo

siinä vaiheessa verkon yli. Salasana ei paljastu ainoastaan seurannan suorittajalle vaan salasana voi altistua myös verkkovakoilulle, mikäli seurantapalvelu lähettää tiedot salaamattoman yhteyden yli.

4 GOOGLE ANALYTICS KÄVIJÄSEURANTA

4.1 Google Analytics käyttöliittymän raportit

Google Analytics -verkkosivusto tarjoaa monia eri tapoja käsitellä tallennettua dataa. Palvelun käyttöliittymä on jaettu pääosin viiteen eri raportointityyppiin. Nämä ovat reaaliaikainen seuranta, yleisö, hankinta, käyttäytyminen ja konversiot.

Reaaliaikainen seuranta raportoi sivuston käytön reaaliajassa noin viimeisen 30 minuutin ajalta. 30 minuuttia onkin juuri aika, jolla istunnot määritetään Analyticsissa [33]. Raportista nähdään aktiivisten vierailijoiden määrä, vierailijoiden sijainnit sivustolla, sekä arviot vierailijoiden maantieteellisistä sijainneista.

Yleisö-raportissa raportoidaan halutulta ajanjaksolta keskeisimmät web-analytiikan metriikat, kuten istuntojen määrä, välittömät poistumiset (engl. bounce rate), tai uusien käyttäjien suhde palaaviin käyttäjiin. Viimeksi mainittu metriikka saavutetaankin JavaScriptin avulla asettamalla evästeitä käyttäjien koneille käyttäjien yksilöimiseksi. Oletuksena raportti antaa tiedot kuukauden ajalta ja raportin pääasiallinen käyttö onkin tarkastella kerättyjä tietoja, kuten istuntojen tai käyttäjien määriä, jokseenkin pitkällä aikavälillä trendien löytämiseksi.

Kolmannessa raporttityypissä, eli hankinta-raportissa, on selvitettynä miten käyttäjät on hankittu seurattavalle sivustolle. Siinä luetellaan onko käyttäjä saapunut esimerkiksi hakukoneen kautta tai jonkun toisen sivuston linkittämänä. Nämä näkymät antamat myös kuvan verkkomainonnan, sosiaalisen median, sähköposti- tai muiden kampanjojen synnyttämistä kävijävirroista mikäli mainitut palvelut ovat liitetty Google Analytics seurantaan.

Käyttäytyminen-raportointi kertoo kuinka sivuston kävijät ovat navigoineet sivulla ja mitkä sivut ovat olleet suosituimmat käyttäjien keskuudessa. Kävijän kulku voi paljastaa sivustolla olevia tärkeitä resursseja, joihin kävijät eivät esimerkiksi osaa navigoida. Tämä

raportointityökalu voi olla hyvin tärkeä sivuston sisällöntuottajille ja kehittäjille, sillä se antaa kokonaiskuvan sivuston käytettävyydestä.

Käyttäytyminen-raportti sisältää myös Analytics-tapahtumat. Analytics-tapahtumat mahdollistavat käyttäjien toimintojen seuraamisen erillään sivujen tai näkymien lataamisesta. [18] Ne ovat siis lisätoiminnallisuutta verkkosivun varsinaiseen seurantaan tarkoitettuihin ominaisuuksiin, kuten sivun latauksien seurantaan. Analytics-tapahtumalla on kategoria, jokin sitä kuvaava toiminto ja tapahtuman yksilöivä tunniste. Toiminnot ovat jaettu kategorioihin ja toimintoja voidaan tarkastella osana kategoriaa tai ilman. Verkkosivuston haltija voi määrittää sivulla esiintyville tapahtumille, kuten ostoksen tekeminen tai jonkun elementin klikkaus, erityisen arvon. Tapahtuman arvolla ei ole yksikköä, vaikka arvot voivatkin tulla oikeista luvuista, joille on määritelty yksikkö, kuten valuutat.

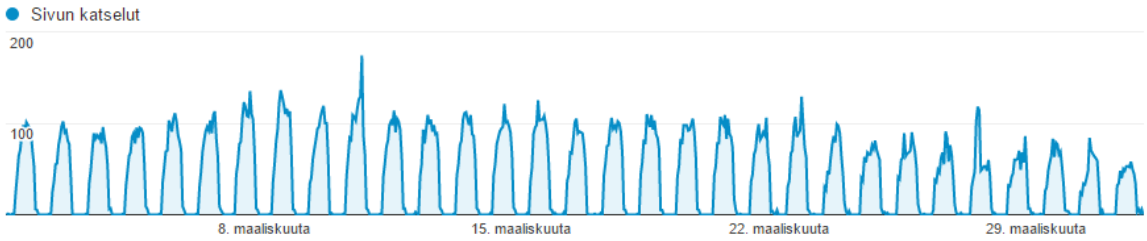
Viimeinen yleinen koko sivuston kattava raporttityyppi on konversiot-raportti. Tämä on tärkeä varsinkin verkkokaupoille, sekä muille kaupallisille sivustoille. Konversioiden toteutumisen seuraaminen on tärkeää, koska pelkkä sivun katselu ei tarkoita, että kyseisen sivun tarkoitus olisi toteutunut [13]. Verkkosivustolla kävijä on saattanut esimerkiksi olla löytämättä sivulta etsimäänsä toimintoa tai tietoa. Konversioiden käyttöönotto vaatii tavoitteiden asettamisen verkkosivustolle. Tavoitteita voi asettaa kahdella tavalla: käyttämällä ohjelmallista käyttöliittymää tai luomalla tavoitteen toteutumista kuvaavan sivun. Tällainen sivu voi olla esimerkiksi rekisteröitymisen jälkeen näytettävä erillinen ”Kiitos rekisteröitymisestä!” -sivu. Käytännössä tämän ominaisuuden käyttöönotto, kuten Analytics-tapahtumat, vaatii suunnitelmallisuutta jo sivuston toteutusvaiheessa.

4.2 Lappeenrannan matkailun i+-päätteiden seuranta

Lappeenrannan matkailun visitlappeenranta.fi-sivustossa Google Analytics -integraatio on toteutettu siten, että data i+-päätteiltä kerätään muusta sivustosta erilliseen Analytics-tiliin. Täten voidaan seurata erikseen päätteille tarkoitettua kioskisivua ja selainkäyttäjille tarkoitettua sivustoa. Tässä tilissä seuranta on toteutettu eri tavalla, käyttäen ainoastaan Analytics-tapahtumia, joiden toimintaa kuvattiin aiemmin.

i+ -toteutuksessa tapahtumat on jaettu käyttöliittymän elementtien mukaan eri kategorioihin. Seurattavia tapahtumakategorioita ovat kielen ja välilehden vaihtaminen, kategorian valitseminen (kategoria viittaa tässä yhteydessä Lappeenrannan tapahtumien kategorisointiin), yrityksen sekä tapahtuman katsominen. Jokainen näistä tapahtumakategorioista voidaan käsitellä erikseen Analytics-käyttöliittymässä. Seurattaville tapahtumille ei ole asetettu arvoja. Arvojen asettaminen ei olisikaan kovin mielekästä, sillä voi olla hankalaa arvottaa jonkin välilehden klikkaaminen toisesta. Toisaalta, on mietittävä, onko tapahtumien käyttö tähän tarkoitukseen perusteltua? Tapahtumat ovat kuitenkin vain lisäominaisuus varsinaiselle sivun katselujen seuraamiselle.

Päätteet suorittavat automaattista sivun päivitystä ajoittain päivittääkseen näkymän edellisen käyttäjän jäljiltä. Nämä päivitykset päätyvät suoraan myös tähän eriytettyyn Analytics-tilissä tapahtuvaan seurantaan sivun katseluina. Tämä on ongelma sillä se tekee suuren osan Google Analyticsin tarjoamista ominaisuuksista hyödyttömiä sillä data on vääristynyt sivun päivitysten takia.

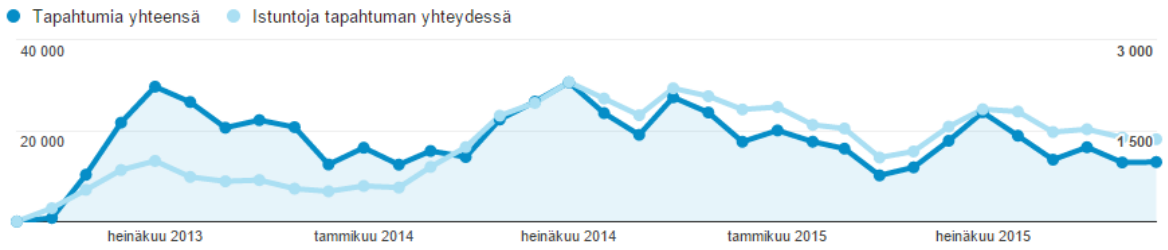


Kuva 4: Sivun katselut i+-kioskisivustoon maaliskuulta 2014 [Google Analytics]

Esimerkiksi kuvassa x nähdään yksi Analyticsin kuvaajista: Yleisö-raportin sivun katselut. Tässä raportissa ei voida tietää automaattisten sivun päivitysten todellista vaikutusta raportin esittämiin tuloksiin. Tämän ja muun kuin i+-päätteiden aiheuttamaa liikennettä pohditaan tarkemmin luvussa 5. Kuvaajan perusteella on kuitenkin selvää, että päivällä tapahtuva automaattinen sivunpäivitys vääristää tuloksia.

Automaattisen sivunpäivityksen takia päätteiden käytön seuraamiseen voidaan käyttää ainoastaan pientä osaa Google Analytics ominaisuuksista eli Analytics-tapahtumia, sillä

sivun latausta ei rekisteröidä tapahtumana Analytics-palveluun. Tämän vuoksi päätteiden tekemillä automaattisilla verkkosivun uudelleenlatauksilla ei ole häiritsevää vaikutusta Analyticsissa tapahtuvaan seurantaan tapahtumien osalta.



Kuva 5: i+-päätteiden käytön kehitys vuosina 2013–2015 [Google Analytics]

Pelkkiä tapahtumiakin tarkastellessa saadaan hyödyllistä tietoa päätteiden käytöstä. i+-päätteiden tapauksessa voidaan istuntojen katsoa olevan käyttäjiä, sillä koneet ovat julkisessa käytössä. Kolmen vuoden tarkastelujaksolta (Kuva 5) huomataan kuinka kuukausittaiset istuntojen määrien huiput ajoittuvat jokaisen vuoden heinäkuulle. Pitkän aikavälin trendiä etsiessä onkin hyödyllistä tarkastella raporttia kuukausittaisella tarkkuudella, sillä päivätasolla päätteiden käyttö vaihtelee suuresti.

Käyttäjien määrä (istunnot) on noussut selkeästi vuoden 2014 aikana. Helmikuussa istuntoja oli 562. Heinäkuulle istuntojen määrä oli noussut huomattavasti: 2 299 istuntoon. Tapahtumien määrää tarkasteltaessa huomataan samankaltainen kausiluonteisuus, mutta pitkän aikavälin trendi on ollut pikemminkin laskeva. Tapahtumien määrä ja istunnot liittyvät kuitenkin vahvasti toisiinsa. Vuonna 2013 tapahtumia on ollut suhteessa enemmän käyttäjiin verrattuna, joka kertoo keskimääräisen käyttäjän käyttäneen laitetta enemmän. Tämä johdettu mittari, tapahtumien määrä jaettuna istunnoilla, kuvaa siten käyttäjän kiinnostusta ja istunnon kestoa. Google Analyticsissa tällaiset johdetut mittarit voidaan luoda käyttäen *Lasketut tiedot* -ominaisuutta [34], jolloin ei tarvitse vertailla kahta eri mittaria.

4.3 i+-päätteiden tapahtumien tarkastelu kuukauden ajalta

Tarkastelujaksoksi valittiin maaliskuu 2014 samoin kuin saatavilla olevien verkkosivuston palvelimen lokien tarkastelussa myöhemmissä luvuissa. Täten voidaan vertailla osaltaan millaisia eri tuloksia saadaan Google Analytics ja lokien tarkasteluun perustuvilla menetelmillä.

KPI	*Arvo*
Tapahtumia yhteensä	15 529
Yksilöidyt tapahtumat	2 107
Istuntoja tapahtuman yhteydessä	902
Tapahtumat / Istunto, johon sisältyy tapahtuma	17,22

Taulukko 1: Tapahtumien KPI-mittarit

Analytics-tapahtumien pääsivulla näytetään tärkeimmät tapahtumia koskevat mittarit (Taulukko 1). Päätteiden käyttöä tarkasteltaessa tärkein mittari kuvaa istuntoja tapahtuman yhteydessä. Siinä kerrotaan istunnot, joissa on ollut vähintään yksi tapahtuma mukana. Nämä tapahtumat ovat käytännössä käyttäjien tekemiä klikkauksia päätteellä, joten luvusta saadaan realistinen kuva käyttäjistä kuukauden aikana. Istuntojen määrään liittyy olennaisesti istuntojen aikakatkaisun pituus, joka on oletuksena 30 minuuttia. 30 minuuttia voi olla varsin pitkä aika i+-päätteelle, jossa sama sivusto on aina auki ja käyttöaika on lyhyt.

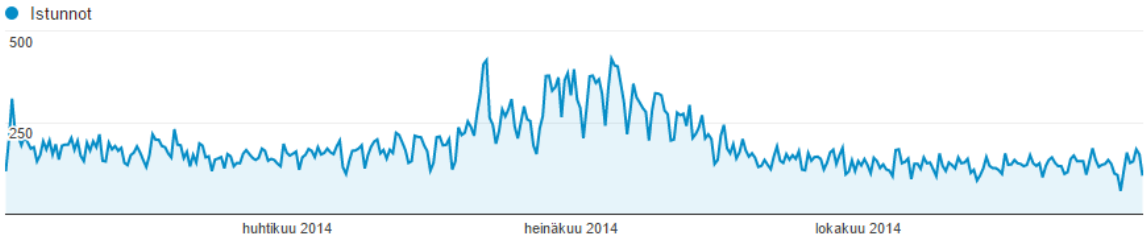
Tapahtumista kieltä vaihdettiin suomeksi 39 %, venäjäksi 38 %, englanniksi 11 %, ja saksaksi 11 % tapauksista. Täytyy kuitenkin huomata, että nämä ovat vain tapahtumia eivätkä kerro, sitä jos käyttöliittymä oli käyttäjän saapuessa jo oikealla kielellä. Sivuston oletuskielenä on suomi, joten suomen kielen käyttö voi olla paljon suurempi kuin raportti antaa ymmärtää.

Kategorian ja välilehden vaihtamiseen liittyvissä tapahtumissa ongelmaksi nousee samankaltaisten tapahtumien pirstoutuminen eri toimintoihin kielen mukaan. Näiden kahden tapahtumakategorian tapahtumia ei tallenneta käyttäen kielestä riippumatonta toiminnon tunnistetta vaan tapahtumat menevät eri toimintojen alle riippuen kielestä vaikka kyseessä olisi täysin sama toiminto. Tämä toteutustapa vaikeuttaa raportin lukemista ja vaatii lukijaa tekemään käsin työtä lukujen yhteen laskemiseksi, sekä käytännössä osaamaan kaikkia kieliä, joita i+ -verkkosivusto tukee.

Suosituin kategoria oli nähtävyydet molemmilla kielillä, sekä suomella että venäjällä. Toisaalta välilehtien tapauksessa, jotka kuvaavat millaista tietoa vierailijat etsivät sivulta, venäjän kieliset käyttäjät valitsivat ostokset-välilehden, kun taas suomen kielellä käyttävät valitsivat useimmin tapahtumat-välilehden. Lappeenrannan tapauksessa lienee ymmärrettävää, että juuri ostokset kiinnostavat venäläisiä kävijöitä.

4.4 Vertailu Cross-Border Travel -verkkosivustoon

Tässä alaluvussa vertaillaan Visit Lappeenranta -sivuston käyttöä Cross-Border Travel -sivustoon vuoden 2014 ajalta. Molemmilla sivustoilla on samankaltainen käyttäjäkunta ja molemmat ovat Google Analytics seurannassa.



Kuva 6: Visit Lappeenranta -verkkosivun istunnot vuonna 2014 [Google Analytics]

Vaikkakin sivustoissa on yhtäläisyyksiä, avautuvat niiden erot monin eri tavoin Analytics-raporteissa. Lappeenrannan matkailun verkkosivuston (ei sisällä i+-päätteiden kioskisivuja) käyttö ajoittuu vahvasti kesäkuukausiin. Kuvassa 6 istuntojen määrissä nähdään selvä piikki kesän aikana. Tällaista trendiä ei ole havaittavissa CBT-verkkosivuston käytöstä, johtuen osaksi käytön seurannan katkoista, mutta ehkä tärkeimpänä sivuston muista ongelmista.

CBT-sivusto on huomattavasti yksinkertaisempi ja vähemmän tunnetumpi kuin Lappeenrannan matkailun verkkosivusto. CBT-sivustossa istuntoja oli Analyticsin mukaan 3 996 vuoden 2014 aikana ja välitön poistuminen varsin korkea 70 prosenttia. Vastaavasti Visit Lappeenranta -sivustolla 68 053 istuntoa, välittömän poistumisen ollen paljon alhaisempi; 40 prosenttia. CBT-verkkosivulle tulleista kävijöistä siis miltei kolme neljästä poistuu heti ensimmäiseltä sivulta. Tämä voi johtua siitä, että sivusto ei vastaa käyttäjän ennakko-odotuksia verkkosivusta. Käyttäjä voi odottaa saapuvansa aivan erilaiselle sivulle, kuin mille päätyi. Tarkasteltaessa istuntojen alkuperämaata, huomataan, että noin 70 prosenttia istunnoista on peräisin muualta kuin oletetun käyttäjäkunnan maista eli Suomesta tai Venäjältä. Monet kävijät ovat esimerkiksi Yhdysvalloista, Intiasta ja Brasiliasta.

CBT-sivustossa palaavien käyttäjien suhde uusiin oli noin 1:9. Visit Lappeenranta -sivustossa puolestaan hieman korkeampi, noin 2:8. Myös hakukoneiden avulla CBT-sivustolle tulleiden käyttäjien suhde on suurempi. CBT-verkkosivulle kävijät siis tulevat useammin hakukonetulosten ohjaamana muualta kuin ehkä halutulta alueelta. Sivuston nimi onkin varsin universaali ja voi hyvin ymmärtää miten amerikkalainen Meksikossa lomailija voisi eksyä kyseiselle verkkosivustolle. CBT-verkkosivustolle voisikin olla hyödyllistä toteuttaa hakukoneoptimointia ja meta-tietojen lisäämistä sivulle kertomaan, että kyseessä on juuri Lappeenrannan ja Pietarin välille sijoittuvan rajamatkailu.

Vuoden 2015 aikana vanha Cross-Border Travel -sivusto otettiin pois käytöstä ja korvattiin yhdellä sivulla, joka kertoo käyttäjille uuden sivun tulevan pian ja jolla voidaan tilata sähköposti-ilmoitus uuden sivun julkaisemisesta. Tämän seurauksena istuntoja oli vain 910 kappaletta ja välitön poistuminen nousi yli 90 prosenttiin kävijöistä. Sivulla sähköposti-ilmoituksen tilaaminen olisi helppo laittaa seurantaan konversio-ominaisuuden avulla. Sähköposti-ilmoituksen tilaaminen vastaisi siis yhtä konversiota eli onnistunutta kiinnostuneen asiakkaan hankintaa.

5 I+ VERKKOSIVUSTON LOKIEN ANALYSOINTI

5.1 i+-verkkosivuston lokitiedot

Lappeenrannan matkailun i+-järjestelmän päätteiden käyttöä tarkasteltiin palvelimen lokien osalta. Lokidataa kerättiin kuukauden ajalta, maaliskuulta 2014. Loki on tallennettu palvelimelle käyttäen yhdistettyä lokiformaattia [35]. Lokia on yhteensä 1 884 524 riviä, noin 545 Mt kokoisessa tekstitiedostossa. Yksittäinen lokitiedoston rivi tarkoittaa yhtä HTTP-pyyntöä palvelimelle.

Lokitiedosto kattaa koko Visit Lappeenranta -sivuston, joten osa riveistä kannattaa jättää pois jo lokia luettaessa, sillä tarkastelun kohteena ovat ainoastaan i+-päätteet. i+-sivusto sijaitsee osoitteessa */iplus/*. Toisin kuin selaimelle tarkoitettu verkkosivusto, päätteet käyttävät ainoastaan sivuja osoitteessa */iplus/kiosk/*. Lisäksi päätteet päivittävät sisällön AJAX-kutsuilla osoitteesta */iplus/kioskajax/*. Näiden eri sivujen suodattaminen koko lokitiedostosta antaa kokonaiskuvan lokiin tallennetuista pyynnöistä.

Kuvaus	*HTTP-pyynnön filtteröinti*	*Referer kentän filtteröinti*	*Rivit*	*%-osuus*	*IP-osoitteet*
Iplus	GET\s/iplus/\s.*		3978	0,21%	1298
Iplus + alisivut	GET\s/iplus/.*\s.*		1332966	70,73%	4543
Kiosk alisivut	(GET|POST)\s/iplus/kiosk/.*\s.*		3786	0,20%	74
Kiosk	(GET|POST)\s/iplus/kiosk\s.*		25806	1,37%	102
Kiosk + kioskajax	(GET|POST)\s/iplus/kiosk.*\s.*		151116	8,02%	141
Kioskajax	(GET|POST)\s/iplus/kioskajax/.*\s.*		121524	6,45%	141
Kioskajax	(GET|POST)\s/iplus/kioskajax/.*\s.*	.*/iplus/kiosk.*	121307	6,44%	105

Taulukko 2: Lokitiedoston suodattaminen säännöllisillä lausekkeilla

Yllä olevasta taulukosta nähdään, että yli 70 prosenttia kaikesta sivuston lokista liittyy i+-järjestelmään. Tämä sisältää i+-päätteiden sekä selainkäyttäjille tarkoitetun verkkosivuston käyttöliittymät i+ -karttanäkymään.

Pelkästään i+-päätteitä koskevan kioskajax-suodatettujen lokitietojen määrä on 6,44 prosenttia lokista. Ne kuvaavat lähinnä käyttäjän toimintoja ja painalluksia päätteen

käyttöliittymässä, sillä ne vastaavat suoraan käyttöliittymäikonien painalluksia näytöllä. Tämän vuoksi analysointi tehdään pääosin tästä datasta.

Pyynnöt kioskajax-resursseihin ovat siis i+-järjestelmän AJAX-kutsujen aiheuttamaa. Tähän lukuun sisältyy luonnollisesti kaikki kutsut i+-päätteiltä, hakukoneroboteilta ja muilta vierailijoilta. Selainkäyttäjille näkyvä versio on eriytetty i+ päätteiden vastaavasta, eikä se käytä AJAX-pyyntöjä sisällön päivittämiseen. Lisäksi i+-sivua ei linkitetä muualta pääsivulta, mikä vähentää siihen suuntautuvaa liikennettä. Ei voida kuitenkaan olettaa, ettei tavallisia selainkävijöitä ohjata esimerkiksi hakukoneiden hakutulosten perusteella suoraan i+-päätteille tarkoitetulle sivulle. Tämän liikenteen voidaan kuitenkin katsoa olevan varsin pieni.

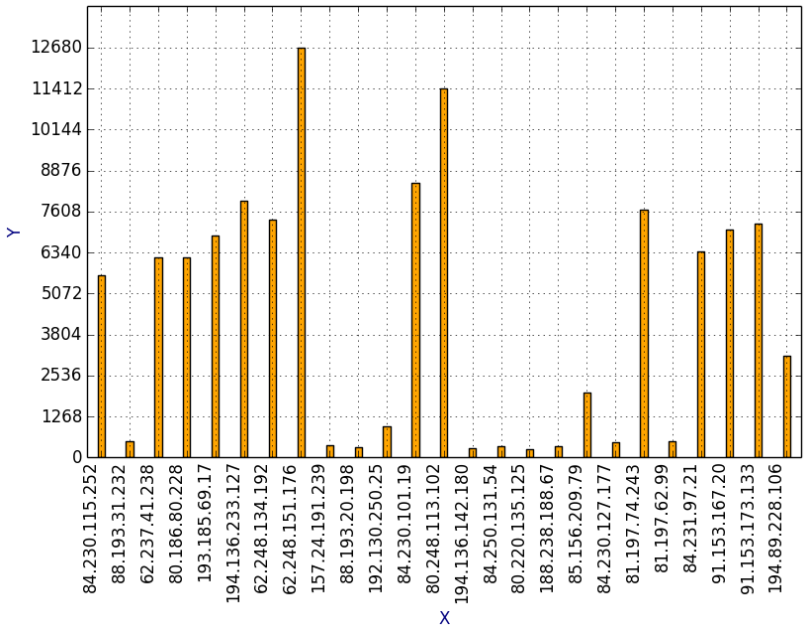
5.2 Python-sovellus lokitietojen analysointiin

Lokitiedoston analysointiin kirjoitettiin yksinkertainen Python 3 -sovellus. Sovellus prosessoi lokitiedoston rivit tietorakenteiksi, jotta lokitiedoston merkintöjä voidaan analysoida. Pyyntöjä ryhmitellään ensiksi IP-osoitteiden mukaan ja sen jälkeen niitä eritellään istunnoiksi käyttämättömänä oloajan mukaan. Käytännössä istunnot, eli sessioidut pyynnöt, kuvaavat monta eri kävijää (sessiota) on kertynyt kuukaudessa kullekin vierailijalle (IP-osoitteelle).

Ohjelma kykenee karsimaan rivejä säännöllisten lausekkeiden avulla jo lukuvaiheessa. Tällä voidaan laskea mukaan ainoastaan pyynnöt, jotka kohdistuvat ainoastaan i+-päätteille tarkoitettuun sivuun. IP-osoitteiden mukaan lajitelluista pyynnöistä tehtiin CSV-raportteja sekä piirrettiin kaavioita matplotlib-kirjaston avulla.

5.3 i+-päätteiden tunnistaminen lokitietueiden perusteella

Analysoitaessa päätteiden AJAX-pyyntöjä huomataan, että sivujen liikenne jakautuu suurimmaksi osaksi 13–15 suurimman IP-osoitteen kesken. Pelkkien IP-osoitteiden perusteella ei kuitenkaan voi vielä tehdä vahvoja johtopäätöksiä, siitä olisiko nämä koneet juuri i+-päätteitä.



Kuva 7: 25 pyyntömääriltään suurinta IP-osoitetta (/iplus/kioskajax)

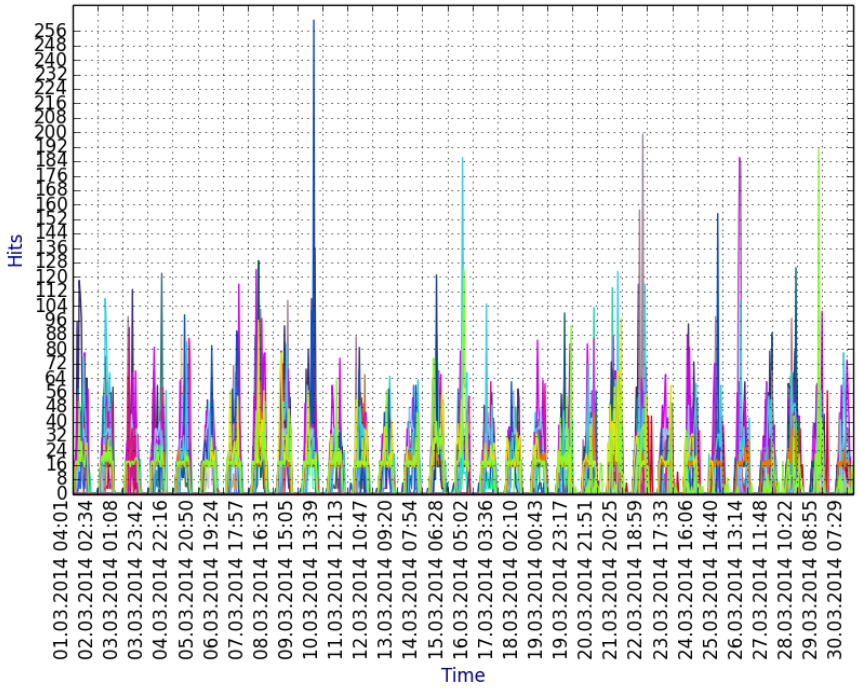
Tarkasteltuamme näiden IP-osoitteiden UAS-informaatiota huomataan, että 15 suurimman pyyntömäärien antamissa koneissa on sama UAS. Kyseinen tietue kertoo koneiden olevan Windows NT 5.1 eli Windows XP -käyttöjärjestelmällä ajettuja Internet Explorer 8.0 -selaimia. Koneet IP-osoitteiden takana käyttävät siis samaa selainta ja käyttöjärjestelmää mitä i+-päätteissä. UAS kertoo lisäksi koneen .NET-tuesta ja huomauttaa selaimen käyttävän Microsoftin Trident-selainmoottoria. UAS ei kuitenkaan ole kovin luotettava tieto yksilöimään päätteitä tai edes selventämään onko kyseessä jokin käytössä olevista päätteistä.

Päätteiden suorittama automaattinen sivun päivitystoiminto voi luoda paljon liikennettä. Sivusto tekee jokaisen uudelleenlatauksen yhteydessä kolme eri AJAX-pyyntöä. 10 minuutin välein tehtävä automaattinen sivun päivitys tarkoittaisi yhteensä siis 18 pyyntöä tunnissa. Käyttäjämäärää voidaan siten arvioida tarkemmin kun tiedetään automaattisen sivunpäivityksen vaikutus tuloksiin.

On hyvin mahdollista, että ainakin osassa i+-päätteitä IP-osoite vaihtuu ajoittain, joka vaikeuttaa koko kuukauden kävijämäärien arviointia. Joka tapauksessa myös dynaamisia IP-osoitteita käyttävät koneet voivat pitää saman osoitteen koko tarkastelujakson ajan. Tällaiset

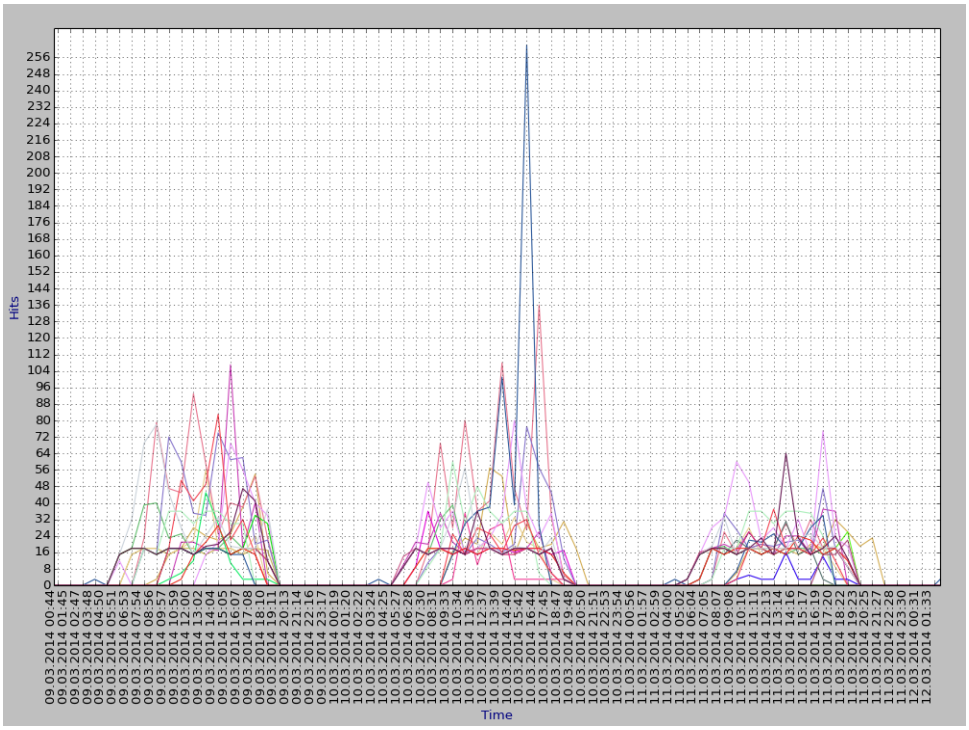
tapaukset ovatkin erityisen mielenkiintoisia tarkastella, jos voidaan osoittaa että kyseinen IP-osoite kuuluu jollekin i+-päätteelle.

5.4 AJAX-pyyntöjen analysointi



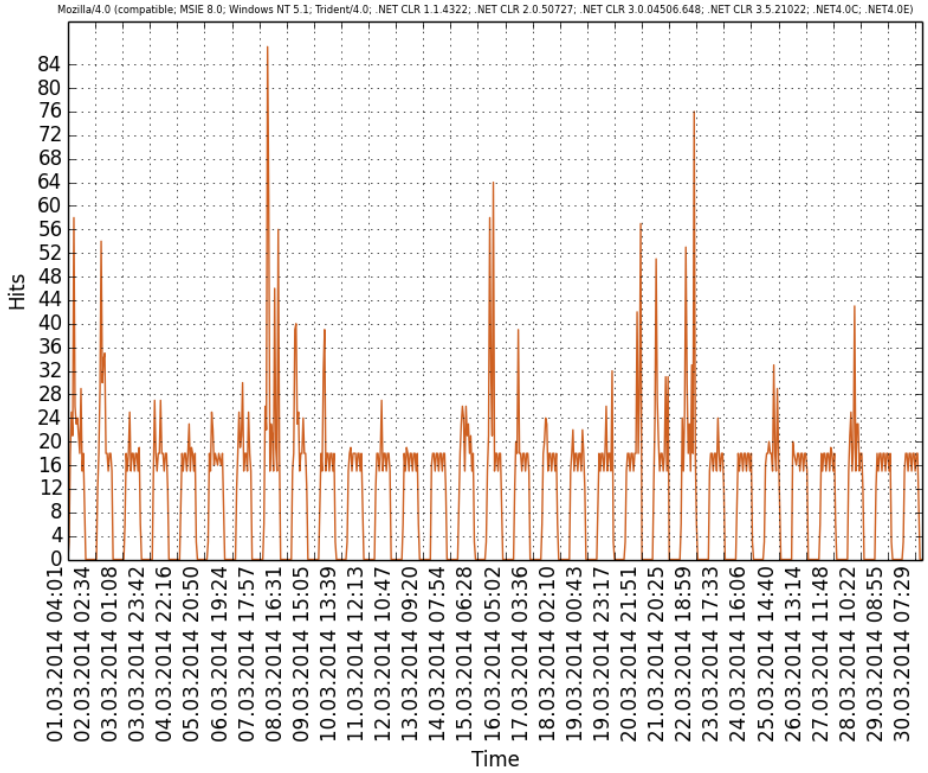
Kuva 8: Pyynnöt tunnin aikana ajan funktiona jokaiselle IP-osoitteelle

Kaaviokuvassa yllä on piirrettynä eri värillä jokaisen IP-osoitteen pyynnöt kioskajax-resursseihin. Kaaviosta näkyy kuinka pyynnöt jaksottuvat maaliskuun kaikille 31 päivälle. Päivittäiset ruuhkahuiput ovat selvästi näkyvissä kaaviossa, huolimatta automaattisesti sivun päivityksestä. Tapauksessa, jossa päätteillä ei olisi käyttöä ja missä ainoa liikennettä aiheuttava tekijä olisi automaattinen sivun päivitystoiminto, olisivat kaavion kuvaajat horisontaalisesti lineaarisia.



Kuva 9: Pyynnöt tunnin aikana ajan funktiona jokaiselle IP-osoitteelle kolmen päivän ajalta

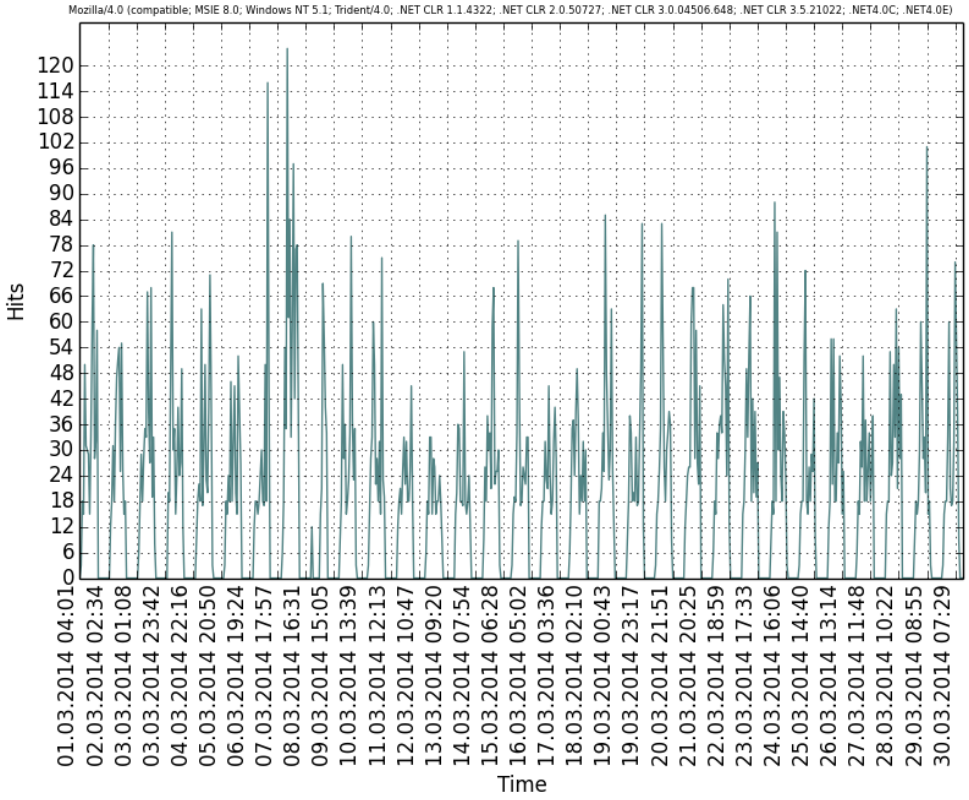
Yksittäisten päivien tarkastelu näyttää, ettei yöaikaan ole juurikaan liikennettä. Automaattinen sivun uudelleenlataus ei siten myöskään aiheuta liikennettä yöaikaan. Kaaviosta voidaan nähdä kuinka käyttö ajoittuu iltapäivään ja välillä saadaan suhteellisen suuria käyttömääriä.



Kuva 10: Mahdollinen pääte (194.136.233.127), jolla on vain ajoittaista käyttöä.

Toisaalta päätteiltä, joilla ei ole paljon käyttöä voidaan nähdä automaattisten sivupäivitysten vaikutus käyttömäärän laskemiseen. Kuvassa 10 on pääte, jolla on käyttöjä vain osassa tarkastelujakson päivistä. Todennäköisesti päätteellä päivitetään sivu päiväsaikaan 10 minuutin välein aiheuttaen yhteensä noin 18 pyyntöä tunnissa myös päivinä jolloin ei ole käyttäjiä.

Huolimatta automaattisen päivityksen tuloksia vääristävästä vaikutuksesta, on mielenkiintoista verrata kuvassa 10 olevaa ajoittain käytettyä mahdollista päätettä usein käyttöä näkevään päätteeseen kuvassa 11.



Kuva 11: Mahdollinen pääte (62.237.41.238), jolla on eniten käyttöä.

Näiden kahden edellä esitetyn päätteen lisäksi saatiin vastaavia kuvaajia kaikille muillekin IP-osoitteille. Muutamille IP-osoitteille on saatavilla tietoja vain osalle päiviä, kuitenkin automaattisen päivityksen vaikutuksen edelleen näkyessä tuloksissa. Tämä kertoo tilanteesta, jossa päätteen IP-osoite on vaihtunut kesken tarkastelujakson. Useimmissa kuvaajissa on kuitenkin näkyvissä pyyntöjä vain yhden päivän ajalta, joka selittää satojen IP-osoitteiden kerääntymisen myös siinä tapauksessa, että suodatetut lokitiedot sisältävät ainoastaan oikeita i+-päätteitä. Näiden tulosten pirstoutumisen takia ei ole mielekästä ryhtyä yhdistämään eri IP-osoitteiden tietoja.

5.5 Sessioidut pyynnöt i+-päätteiltä

Pyyntöjen jako istuntoihin eli sessionti tehtiin siten, että peräkkäiset pyynnöt luetaan kuuluvaksi samaan sessioon, kun niiden välinen aikaleima on pienempi kuin määritelty aika *t*. Taulukossa 3 on listattuna sessionnin tulokset eri *t* arvoilla.

IP-osoite	*sessiot (t=1 min)*	*sessiot (t=3 min)*	*sessiot (t=5 min)*	*sessiot (t=10 min)*	*sessiot (t=15 min)*	*sessiot (t=30 min)*	*sessiot (t=2 h)*	*sessiot (t=24 h)*	*Pyynnöt yhteensä*
62.237.41.238	2693	2361	2132	1701	33	32	32	1	12683
80.248.113.102	2546	2133	1823	1280	34	31	31	1	11424
80.220.135.125	2076	1896	1769	1530	31	30	30	2	8483
194.136.233.127	2351	2279	2219	2109	32	31	31	1	7925
84.250.131.54	1713	1496	1349	1091	31	31	31	1	7669
194.89.228.106	1358	1270	1245	1091	56	52	36	2	7353
91.153.167.20	1693	1474	1334	1048	99	66	29	1	7220
88.193.31.232	1649	1483	1366	1118	142	104	47	1	7054
194.136.142.180	1719	1597	1534	1425	119	100	49	1	6858
88.193.20.198	1961	1924	1902	1859	31	28	28	1	6379
192.130.250.25	1946	1918	1883	1816	33	31	31	1	6196
193.185.69.17	1625	1503	1417	1251	22	21	21	1	6185
157.24.191.239	1449	1361	1312	1207	88	75	50	1	5640
91.153.173.133	576	426	373	306	114	95	46	2	3157
84.230.127.177	602	588	579	555	9	9	9	1	2012
80.186.80.228	93	72	67	58	52	38	11	1	975
85.156.209.79	55	47	45	43	42	38	23	1	511
188.238.188.67	6	2	1	1	1	1	1	1	498
84.231.97.21	154	151	149	144	4	3	3	1	482
62.248.134.192	96	73	62	49	1	1	1	1	390
81.197.74.243	79	73	70	66	1	1	1	1	352
84.230.115.252	82	76	70	61	1	1	1	1	338
62.248.151.176	81	75	71	64	1	1	1	1	305
81.197.62.99	40	24	20	15	12	7	1	1	281
84.230.101.19	76	74	71	68	1	1	1	1	260

Taulukko 3: Tulokset sessioiduista AJAX-pyynnöistä

Automaattisesti tehtävä sivun uudelleenlataus vaikeuttaa luotettavan sessioinnin saavuttamista huomattavasti. Yllä kuvatulla menetelmällä tehty sessionti tuottaa sessioita kolmeen kategoriaan. Vuorokauden mittaisella sessionnilla (t = 24 h) saadaan yleisesti vain yksi sessio eli jokaisella päivällä on edes jotain käyttöä. Uudelleenlatauksen ajanjaksoa suuremmilla t-arvoilla kaikki pyynnöt jaetaan yhteen sessioon jokaiselle kuukauden 31

päivälle. Pienemmillä t-arvoilla sessionti puolestaan pystyy havaitsemaan uusia sessiota, mutta se ei pysty erottamaan uudelleenlatauksia lähelle sijoittuvaa käyttöä.

Taulukossa 3 on korostettuna päätteet, joista on dataa koko tarkastelujakson ajalta. Muissa IP-osoite on todennäköisesti vaihtunut kesken mittausajanjakson ja data on siten hajautunut usean IP-osoitteen alle. Korostetuiden päätteiden data on siten vertailukelpoista keskenään.

Automaattisten uudelleenlatausten vaikutus alle 10 minuutin t-arvon sessioihin on huomattava. Joka päivä tehtävä 10 minuutin välein tehtävä uudelleenpäivitys aiheuttaa karkeasti noin 2000 sessiota kuukaudessa. Vaikka tiedetään automaattisten pyyntöjen määrä suhteellisen tarkasti voi sessioiden sijoittuminen lähekkäin toisia vääristää arviota automaattisten päivitysten aiheuttamista sessioista. Tästä ylimenevä osa on käyttäjien aiheuttamaa, jota voidaan karkeasti arvioida. Kuitenkaan kovin luotettavan arvion antaminen ei ole mahdollista.

6 POHDINTA JA TULEVAISUUS

6.1 Google Analytics integraation parantaminen i+-järjestelmässä

Huomattavaa on, ettei Google Analyticsin integrointia i+ -järjestelmään ole tehty käyttäen sen monia hyödyllisiä ominaisuuksia. Google Analyticsin käyttöönotto on nopeaa, ja parhaassa tapauksessa ohjelmistoteknisesti yksinkertaista. Kuitenkin integraation toteuttaminen tulisi olla suunnitelmallista. Halutut seurannan ominaisuudet tulisi olla jo tiedossa ohjelmiston määrittelyvaiheessa, sillä useat palvelun ominaisuudet, kuten konversiot, voivat vaatia esimerkiksi kokonaan uusien sivujen tuottamisen verkkosivustolle.

Lappeenrannan matkailun tapauksessa, kuten usein uuden ohjelmiston hankinnassa, on ohjelmisto tuotettu alihankintana. Google Analytics -raportointia käyttää ainoastaan verkkosivuston omistaja, eikä verkkosivun pääasiallinen käyttäjäkunta, jolloin tällaisten ei-toiminnallisten vaatimusten määrittely voi olla helppo sivuuttaa kehittäjän toimesta. Ohjelmiston hankkijan olisikin syytä tiedostaa Analyticsin, tai minkä tahansa muun halutun seurantapalvelun tarjoamat ominaisuudet, ja määritellä ne jo ohjelmiston vaatimusmäärittelyssä.

Kioskisivujen seurannan osalta keskittyminen pelkästään tapahtumien keräämiseen on ongelmallista, sillä se ohittaa suuren osan Google Analyticsin tarjoamista ominaisuuksista. Tärkeimpänä ominaisuutena on kerätä sivun katseluita. AJAX-painotteisissa toteutuksissa, kuten i+ kioskisivut, voidaan sivun katseluina pitää näkymien vaihtumista. Esimerkiksi, juuri kategorian muuttaminen ja tapahtuman avaaminen voisi olla sivun katseluita, sillä ilman AJAX-toteutusta nämä tulisivat automaattisesti luetuiksi sivun katseluiksi. Lisäksi, oletuksena päällä olevan sivun ensimmäisen latauksen raportoiminen sivun katseluna voi laittaa pois päältä. Tämä mahdollistaisi automaattisten sivun päivitysten sulkemisen pois häiritsemästä raportointia.

Google Analytics -integraation syventämisen lisäksi matkailupäätteiden seurantaan voitaisiin ottaa käyttöön muita kävijäseurantatyökaluja Analytics-seurannan rinnalle. Esimerkiksi Yandex Metrica tarjoaa WebVisor-palvelun, jolla voidaan seurata kursorin

liikkumista sivulla reaaliajassa. Toisaalta päätteitä voi olla hyödyllistä seurata pelkän lokien analysoinnin avulla, sillä monia Analyticsin tarjoamia ominaisuuksia ei tarvita, päätteiden ollessa ainoat seurattavat laitteet.

6.2 Päätteiden eriyttäminen i+-järjestelmässä

i+ -järjestelmän eräänä seurannan tavoitteena on eriyttää kävijämäärät päätekohtaisesti. Tähän olisi useita mahdollisia ratkaisuja, sekä Google Analytics -ratkaisussa että lokien analysointiin perustuvissa menetelmissä.

Käytössä olevan automaattisen sivunpäivityksen takia tulisi Analytics-istuntoevästeen aikakatkaisua lyhentää oletuksena olevasta 30 minuutista. Tämän pystyy tekemään käyttämällä Analyticsin rajapintaa. [36] Aikakatkaisun lyhentäminen sallisi istuntojen muodostumisen, mutta se ei kuitenkaan poista sivunpäivitysten tuottamaa turhaa liikennettä. Ratkaisu edellä olevaan ongelmaan on olla laskematta sivunlatauksia ollenkaan. Tällöin pelkästään tapahtumat, kuten nappien painallukset, lasketaan. Tämä on täysin hyväksyttävää, varsinkin kun kyseessä on päätetyyppinen ratkaisu, jossa näytetään vain yhtä sivua.

Vaikka istunnot voidaankin huomioida, se ei edelleenkään anna tietoa päätekohtaisesti. Google Analytics ei anna tietoa julki laitteiden IP-osoitteista, tai pyri yleensäkään eriyttämään sivun katseluita laitekohtaisesti. Google Analytics -ominaisuus, *Kustomoidut ulottuvuudet ja mittarit,* kuitenkin antaa mahdollisuuden segmentoida liikennettä ja mitata suhteita näiden segmenttien välillä. Segmentit voisivat i+ -päätteiden tapauksessa olla itse päätteitä ja päätteen sijoituspaikka. Tämä ominaisuus tunnettiin aiemmassa Analytics-versiossa nimellä *kustomoidut muuttujat*. Käytännössä jokainen pääte lähettäisi tapahtumaraportoinnin yhteydessä laitetunnuksen tai paikan nimen.

Lokaalisti laitetunnus voidaan tallentaa muuan muassa evästeeseen, tai uudemmissa selaimissa saatavilla olevaan HTML paikalliseen muistiin. i+-päätteet ovat etäältä hallittavia koneita, joten voi olla helpointa luoda rajoitettu sivu koneen rekisteröimiseen, joka tallentaa laitetunnus-evästeen koneelle. Toinen vaihtoehto voisi olla laitetunnuksen sijoittaminen

URL-osoitteeseen, josta sivun JavaScript voi tunnuksen hakea. Jokaisella päätteellä olisi siis oma URL, josta ne i+-sivun hakevat.

Lisäyksenä havaittuihin ongelmiin Analytics-tapahtumien pirstoutumisessa, kustomoituja ulottuvuuksia ja mittareita voisi käyttää tällä hetkellä pelkkien tapahtumien tallennuksen sijasta myös niiden kategorisointiin sivun sisällä. Sivuston käyttöliittymä on jaettu useisiin eri kategorioihin ja niiden tarkastelu edellä mainitulla ominaisuudella voisi tuottaa helpommin vertailtavaa tietoa ja raportteja.

6.3 Lokien analysoinnin helpottaminen i+-järjestelmässä

Luvussa 5 analysoitiin verkkosivuston lokia kuukauden ajalta ja pyrittiin selvittämään mitä voidaan oppia i+ -päätteiden käytöstä lokitietojen avulla. Varsinkin päätteiden yksilöiminen olisi tärkeää päätekohtaisten kävijämäärien selvittämisessä. Tällä hetkellä ainoa mahdollisuus on turvautua dynaamiseen IP osoitteeseen, UAS-tietueeseen, sekä pyydettyjen resurssien suodattamiseen. Mikään näistä ei ymmärrettävästi tuota kovin luotettavaa ja yksilöivää tietoa, sillä IP-osoite voi muuttua kesken tarkastelujakson ja UAS voi olla sama eri koneilla. Tietoa ei ole tarjolla tarpeeksi eli entropia on liian vähäinen, jotta voitaisiin muodostaa luotettava sormenjälki, kuten luvussa 3.4 käsitellyissä menetelmissä.

Kontrolloidussa järjestelmässä, jossa tunnistettavat päätteet ovat hallittavissa, yksinkertainen ratkaisu ongelmaan olisi tarjota päätteen tunniste sivun ja AJAX -kutsujen pyynnöissä mukana. Tämä mahdollistaisi laitekohtaisen tiedon tallentamisen lokeihin. Pääte voisi tunnistautua käyttäen HTTP BA -menetelmää (Basic Authentication), joka on yksinkertainen Base64-koodattu tunnuksen ja salasanan yhdistelmä lähetettynä pyynnön mukana Authorization-otsikkokentässä. Yleisimmissä Apache 2 lokiformaateissa BA-menetelmän tunnus päätyy suoraan lokiin [35]. Käytännössä salasanaa ei edes tarvita sillä tunnistautumista käytetään ainoastaan kyseisen laitteen aiheuttaman liikenteen tarkasteluun. Luonnollisesti, jos halutaan varmistua etteivät muut tahot pääse vaikuttamaan seurantatuloksiin, pitäisi asettaa vahva salasana. Lisäksi pitäisi käyttää ainoastaan salattua HTTPS-yhteyttä sillä muuten BA-tunnus altistuu muiden pyynnön otsikkotietojen tavoin verkkoliikenteen urkinnalle.

Mikäli halutaan ainoastaan välittää päätteen tunnus lokiin, eikä tarvita salasanalla tunnistautumista, on kyseessä ainoastaan yksinkertainen otsikkotiedon lisääminen AJAX-pyyntöihin. Se, että tunnistetaan ainoastaan klikkaukset AJAX-pyyntöjen kautta tarkoittaa, että lokien analysoinnilla päästäisiin samanlaisiin tuloksiin kuin nykyisiä Google Analytics -tapahtumia käyttämällä. Automaattiset päivitykset eivät siis häiritse tässä tapauksessa enää analysointia, sillä ne voidaan suodattaa pois identiteetti-tiedon niistä puuttuessa. Toisaalta, valmiin ja sopivan, raportteja generoivan, työkalun löytäminen tai toteuttaminen lokien analysointiin voi olla vaikeampaa kuin vain Google Analytics-palvelun käyttäminen.

Eräänä mahdollisuutena olisi käyttää Ident-protokollaa, jossa käyttäjän tunnus päätyy myös lokiin. Toisin kuin BA, jossa tunnus syötetään palvelimelle pyynnön mukana, Ident-protokollassa palvelimen täytyy erikseen ottaa yhteyttä käyttäjän tunnistamiseksi pyynnön jälkeen [37]. Tämä vaatisi kahdensuuntaisen liikenteen ja Ident-palvelimen jokaisella päätteellä, eikä siten ole välttämättä kovin käyttökelpoinen menetelmä kuin tietyissä erityistapauksissa.

6.4 Tiedonlouhinnan työkalujen käytön tutkimus

Luvussa 5 tehdyssä analyysissä kirjoitettiin yksinkertainen Python-työkalu lokien lukemiseen. Työkalun kirjoitusvaiheessa nousi esiin useita käytännön ongelmia, kuten tehokas muistin käyttö suuria tekstitiedostoja lukiessa tai useiden erilaisten filttereiden ja raporttien generoinnin helpottaminen. Käytännössä siis työkalua pitää pystyä konfiguroimaan itse koodista erillään. Työkalun tulisi myös pystyä siirtämään analysoitavaa tietoa kiintolevyn ja muistin välillä ohjelman suorittamisen aikana sekä jaksottamaan tietojen lukua. 32-bittisen Windows-sovelluksen rajat muistille tulevat suhteellisen nopeasti vastaan suuria lokitiedostoja lukiessa.

Saatavilla on kuitenkin myös varsinaisia lokianalysointityökaluja, kuten vapaan lähdekoodin AWStats ja Piwik, joiden käyttöön ei tässä työssä tutustuttu tarkemmin. Näiden työkalujen hyödyllisyys ja ominaisuudet voivat olla mielenkiintoinen tarkastelun kohde, varsinkin vertailtaessa tässä työssä tehdyn lokianalysointityökalun suodatusominaisuuksiin. Samoin

myös syvemmälle matematiikkaan menevillä tiedonlouhinnan menetelmillä, kuten verkkoteorian soveltamisella, voi olla mielenkiintoisia käyttökohteita lokien analysoinnissa.

LÄHDELUETTELO

[1] Hasan, L., Morris, A., & Probets, S. (2009). Using Google Analytics to Evaluate the Usability of E- Commerce Sites. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5619.

[2] Zheng, G., & Peltsverger, S. (2015). Web Analytics Overview. Encyclopedia of Information Science and Technology, Third Edition, (JANUARY 2014), 7674–7683. doi:10.4018/978-1-4666-5888-2.ch756.

[3] Ajantasainen lainsäädäntö. Tietoyhteiskuntakaari. 7.11.2014/917, Luku 23, 205 §. [verkkodokumentti]. Finlex. [Viitattu 21.11.2015]. Saatavissa: http://www.finlex.fi/fi/laki/ajantasa/2014/20140917.

[4] Plaza, B. (2011). Google Analytics for measuring website performance. Tourism Management, 32(3), 477–481. doi:10.1016/j.tourman.2010.03.015.

[5] Norguet, J., Zim, E., & Steinberger, R. (2006). Improving Web Sites with Web Usage Mining , Web Content Mining , and Semantic Analysis. Sites The Journal Of 20Th Century Contemporary French Studies, 430–439. doi:10.1007/11611257_41.

[6] Matheson, M., Martin, P., Lo, J., Ng, J., Tan, D., & Thomson, B. (2013). Intelligence for the personal web. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7855 LNCS, 11.

[7] Nakatani, K., & Chuang, T. (2011). A web analytics tool selection method: an analytical hierarchy process approach. Internet Research, 21(2), 171–186. doi:10.1108/10662241111123757.

[8] Tahir, A., McArdle, G., & Bertolotto, M. (2011). A web-based visualisation tool for analysing mouse movements to support map personalisation. Database Systems for Adanced …, 132–143. doi:10.1007/978-3-642-20244-5_13.

[9] Wang, X., Shen, D., Chen, H., & Wedman, L. (2011). Applying web analytics in a K-12 resource inventory. The Electronic Library, 29(2006), 20–35. doi:10.1108/02640471111111415.

[10] Fagan, J. C. (2014). The suitability of web analytics key performance indicators in the academic library environment. Journal of Academic Librarianship, 40(1), 25–34. doi:10.1016/j.acalib.2013.06.005.

[11] Phippen, A., Sheppard, L., Furnell, S., Phippen, a., Sheppard, L., & Furnell, S. (2004). A practical evaluation of Web analytics. Internet Research, 14(4), 284–293. doi:10.1108/10662240410555306.

[12] Pakkala, H., Presser, K., & Christensen, T. (2012). Using Google Analytics to measure visitor statistics: The case of food composition websites. International Journal of Information Management, 32(6), 504–512. doi:10.1016/j.ijinfomgt.2012.04.008.

[13] Kent, M. L., Carr, B. J., Husted, R. A., & Pop, R. A. (2011). Learning web analytics: A tool for strategic communication. Public Relations Review, 37(5), 536–543. doi:10.1016/j.pubrev.2011.09.011.

[14] Boda, K., Foeldes, A. M., Gulyas, G. G., & Imre, S. (2012). User Tracking on the Web via Cross-Browser Fingerprinting. Information Security Technology for Applications, 7161, 31–46. Retrieved from <Go to ISI>://WOS:000310342000004.

[15] Phippen, a. D. (2004). An evaluative methodology for virtual communities using web analytics. Campus-Wide Information Systems, 21(5), 179–184. doi:10.1108/10650740410567518.

[16] i+ matkailuneuvontalaitteiden sijaintipaikat. [verkkosivu]. Visit Lappeenranta. [Viitattu 23.11.2015]. Saatavissa: http://www.visitlappeenranta.fi/iplus-sijainti.

[17] Referenssit. [verkkosivu]. Mainostoimisto MediaNyt. [Viitattu 23.11.2015]. Saatavissa: http://www.medianyt.fi/referenssit.php.

[18] Event Tracking. [verkkosivu]. Google Developers. [Viitattu 23.11.2015]. Saatavissa: https://developers.google.com/analytics/devguides/collection/analyticsjs/events.

[19] i+ verkkosivu. [verkkosivu]. Visit Lappeenranta. [Viitattu 2.2.2016]. Saatavissa: http://www.visitlappeenranta.fi/iplus/kiosk.

[20] Cross-Border Travel [verkkosivusto]. [Viitattu 2.10.2013]. Saatavissa: http://www.crossbordertravel.eu.

[21] Tracking Code Overview. [verkkodokumentti]. Google. [Viitattu 2.10.2013]. Saatavissa: https://developers.google.com/analytics/resources/concepts/gaConceptsTrackingOverview.

[22] Server Log Analytics. [verkkosivu]. Piwik.org. [Viitattu 2.10.2013]. Saatavissa: http://piwik.org/log-analytics/.

[23] Panopticlick. [verkkosivu]. Electronic Frontier Foundation. [Viitattu 13.11.2013]. Saatavissa: https://panopticlick.eff.org/.

[24] Fourie, I., & Bothma, T. (2007). Information seeking: an overview of web tracking and the criteria for tracking software. Aslib Proceedings, 59(3), 264–284. doi:10.1108/00012530710752052.

[25] Uniform Resource Locators (URL). [verkkodokumentti]. The Internet Engineering Task Force. [Viitattu 2.2.2016]. Saatavissa: https://tools.ietf.org/html/rfc1738.

[26] Custom campaigns. [verkkosivu]. Google. [Viitattu 2.2.2015]. Saatavissa: https://support.google.com/analytics/answer/1033863.

[27] Google Analytics Cookie Usage on Websites. [verkkosivu]. Google Developers. [Viitattu 2.2.2015]. Saatavissa: https://developers.google.com/analytics/devguides/collection/analyticsjs/cookie-usage.

[28] Aho, M. Pari sanaa mittareista. [verkkoartikkeli]. Rongo Oy. 5.6.2012. [Viitattu 20.11.2015]. Saatavissa: http://www.rongo.fi/2012/06/pari-sanaa-mittareista/.

[29] Cookies. [verkkodokumentti]. Euroopan komissio. EU. [Viitattu 2.2.2016]. Saatavissa: http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm.

[30] Do Not Track. [verkkodokumentti]. Firefox, Mozilla. [Viitattu 2.2.2016]. Saatavissa: https://www.mozilla.org/en-US/firefox/dnt/.

[31] IP Anonymization in Google Analytics. [verkkodokumentti]. Google. [Viitattu 2.2.2016]. Saatavissa: https://support.google.com/analytics/answer/2763052?hl=en.

[32] Guidelines for Hamburg-based website operators using Google Analytics. [verkkodokumentti]. HmbBfDI. [Viitattu 2.2.2016]. Saatavissa: https://www.datenschutz-hamburg.de/uploads/media/GoogleAnalytics_Guidelines_for_Hamburg_01.pdf.

[33] How a session is defined in Analytics. [verkkodokumentti]. Google. [Viitattu 2.2.2016]. Saatavissa: https://support.google.com/analytics/answer/2731565?hl=en&ref_topic=1012046.

[34] About calculated metrics [beta]. [verkkosivu]. Google. [Viitattu 22.2.2014]. Saatavissa: https://support.google.com/analytics/answer/6121409?hl=en.

[35] Log Files. [verkkodokumentti]. The Apache Software Foundation. [Viitattu 2.2.2016]. Saatavissa: https://httpd.apache.org/docs/2.2/logs.html.

[36] Tracking Code: Basic Configuration. [verkkodokumentti]. Google Developers. [Viitattu 2.2.2016]. Saatavissa: https://developers.google.com/analytics/devguides/collection/gajs/methods/gaJSApiBasicConfiguration.

[37] Apache Module mod_ident. [verkkodokumentti]. The Apache Software Foundation. [Viitattu 2.2.2016]. Saatavissa: https://httpd.apache.org/docs/2.2/mod/mod_ident.html.

[38] Yksityiskohtainen historia, Yritys. [verkkosivu]. Google. [Viitattu 2.2.2016]. Saatavissa: https://www.google.com/about/company/history/.

LIITE 1. Tyypillinen HTTP-pyyntö

```
GET /
Host: www.example.net
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:25.0) Gecko/20100101
Firefox/25.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: fi-fi,fi;q=0.8,en-us;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate
Cookie: _ga=GA1.2.646592288.1383110865; _ym_visorc=w
Referer: http://www.othersite.net/page/
DNT: 1
Connection: keep-alive
```

LIITE 2. Google Analytics HTTP-pyyntö-vastaus (ga.js, Firefox 25.0, Firebug 1.12.4, http://www.crossbordertravel.eu)

Pyyntö:

```
GET
/__utm.gif?utmwv=5.4.5&utms=1&utmn=1730107934&utmhn=www.crossbordertravel
.eu&utmcs=UTF-8&utmsr=1920x1080&utmvp=1710x603&utmsc=24-bit&utmul=fi-
fi&utmje=1&utmfl=11.9%20r900&utmdt=Cross-
Border%20Travel%20%7C%20Problem%20free%20traveling%20across%20the%20Finni
sh-Russian%20border&utmhid=1139298331&utmr=-
&utmp=%2F&utmht=1384534483272&utmac=UA-30551435-
1&utmcc=__utma%3D88779209.1765434091.1384534483.1384534483.1384534483.1%3
B%2B__utmz%3D88779209.1384534483.1.1.utmcsr%3D(direct)%7Cutmccn%3D(direct
)%7Cutmcmd%3D(none)%3B&utmu=q~ HTTP/1.1
Host: ssl.google-analytics.com
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:25.0) Gecko/20100101
Firefox/25.0
Accept: image/png,image/*;q=0.8,*/*;q=0.5
Accept-Language: fi-fi,fi;q=0.8,en-us;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate
Referer: http://www.crossbordertravel.eu/
Connection: keep-alive
```

Vastaus:

```
HTTP/1.1 200 OK
Age: 88811
alternate-protocol: 443:quic
Cache-Control: private, no-cache, no-cache=Set-Cookie, proxy-revalidate
Content-Length: 35
Content-Type: image/gif
Date: Thu, 14 Nov 2013 16:14:31 GMT
Expires: Wed, 19 Apr 2000 11:43:00 GMT
Last-Modified: Wed, 21 Jan 2004 19:51:30 GMT
Pragma: no-cache
Server: Golfe2
X-Content-Type-Options: nosniff
X-Firefox-Spdy: 3
```

LIITE 3. Google Analytics HTTP-pyyntö-vastaus (analytics.js, Firefox 25.0, Firebug 1.12.4, esimerkki)

Pyyntö:

```
GET
/collect?v=1&_v=j14&a=1933359157&t=pageview&_s=1&dl=http%3A%2F%2Fwww.exam
ple.net&ul=fi-fi&de=UTF-8&dt=Title&sd=24-
bit&sr=1920x1080&vp=1710x603&je=1&fl=11.9%20r900&_u=MAC~&cid=1431565180.1
383935979&tid=UA-43101149-1&z=681546633 HTTP/1.1
Host: ssl.google-analytics.com
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:25.0) Gecko/20100101
Firefox/25.0
Accept: image/png,image/*;q=0.8,*/*;q=0.5
Accept-Language: fi-fi,fi;q=0.8,en-us;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate
Referer: http://www.example.net/page/
Connection: keep-alive
```

Vastaus:

```
HTTP/1.1 200 OK
Access-Control-Allow-Origin: *
Age: 86977
alternate-protocol: 443:quic
Cache-Control: private, no-cache, no-cache=Set-Cookie, proxy-revalidate
Content-Length: 35
Content-Type: image/gif
Date: Thu, 14 Nov 2013 16:27:01 GMT
Expires: Mon, 07 Aug 1995 23:30:00 GMT
Last-Modified: Sun, 17 May 1998 03:00:00 GMT
Pragma: no-cache
Server: Golfe2
X-Content-Type-Options: nosniff
X-Firefox-Spdy: 3
```