Thomas Koinig, BSc

Tracking User Behavior through Web Analytics: An Implementation on the Catrobat Sharing Platform

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Software Development and Business Management

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr.techn. Wolfgang Slany

Co-Supervisor
Dipl.-Ing. Matthias Müller, BSc

Institute of Software Technology

Graz, May 2019

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

The past decades brought a revolution in the World Wide Web (WWW) technology. The Internet protocol did evolve from a stateless protocol, for the transmission of text content, to a transaction protocol for highly interactive web applications. Dr. Stephen Turner introduced the log file analysis to data mine information from the gathered web server requests to discover user statistics. However, the log file analysis soon reached its limits in recording users' client-side interactive interactions. As a consequence, the client-side page tagging was introduced to gather interactions from the user's perspective and has become the most widely used data collection method.

Web analytics gathers data for the optimization of the website design, streamline workflows, adapt the website layout for the audience, or implement functions specific for visitor needs. This thesis introduces web metrics and reports to analyze the visitor behavior and the visitor path through the website.

The use-case scenario for the implementation of web analytics is the Catrobat sharing website. This thesis introduces the trackable content through the page tagging data gathering method. Furthermore, web metrics and reports are introduced to monitor the user behavior of the audience. These web metrics are combined with the customer journey to analyze the user journey of the website visitors.

Kurzbeschreibung

Die World Wide Web (WWW) Technologie hat sich in den vergangenen Jahrzehnte revolutioniert. Das Internet Protokoll entwickelte sich von einem zustandslosen Protokoll für die Übertragung von einfachen Textinhalten zu einem Protokoll für die Übertragung von hoch interaktiven Webanwendungen. Dr. Stephen Turner führte dazu Logdatei Analyse ein um die gesammelte Daten, die durch die Web Server Anfragen entstanden sind zu analysieren. Diese Daten wurden genutzt um das Benutzerverhalten zu erkennen. Allerdings erreichte die Logdatei Analyse bald ihre Limitationen in der Erfassung von Client-seitigen interaktiven Interaktionen. Als Konsequenz wurde zur Erfassung der Interaktionen aus der Benutzerperspektive, das Client-seitige Page Tagging eingeführt, das zur häufigsten eingesetzten Datenerfassungsmethode wurde.

Web Analytics erfasst Daten für der Optimierung des Webseiten Designs, die Vereinfachung von Arbeitsflüssen, die Anpassung des Webseiten Layouts an die Benutzerbedürfnisse, oder die Einführung von speziell an die Besucher angepasster Funktionen. Im Rahmen dieser Masterarbeit werden Web Metriken und Reporte vorgestellt um das Benutzerverhalten und den Navigationspfad durch die Webseite zu analysieren.

Als Anwendungsszenario für die Implementierung von Web Analytics wird die Catrobat Sharing Webseite genutzt. Im Laufe dieser Arbeit wird der überwachbare Seiteninhalt der Webseite durch das Client-seitige Page Tagging vorgestellt. Des Weiteren werden die Web Metriken und Reporte vorgestellt um das Nutzungsverhalten der Besucher erfassen zu können. Im Anschluss werden diese Web Metriken mit der Customer Journey kombiniert um das Nutzungsverhalten der Webseite zu analysieren.

Contents

List of Figures

List of Tables

Listings

1. Glossary

1.1. Abbreviations

ARPANET	Advanced Research Project Agency Network
RIA	Rich Internet Application
KPI	Key Performance Indicator
PI	Performance Indicator
DSS	Decision Support Systems
BI	Business Intelligence
BI&A	Business Intelligence & Analytics
MSS	Management Support Systems
OLAP	Online Analytical Processing
IOT	Internet of Things
HCI	Human-Computer Interaction
MIS	Management Information Systems
EIS	Executive Information Systems
WAA	Web Analytics Association
SAAS	Software as a Service
CMS	Content Management System
CTR	Click-Through-Rate
SEO	Search Engine Optimization
AJAX	Asynchronous JavaScript and XML
HTTP	Hypertext Transfer Protocol
CSS	Cascading Style Sheets
NAT	Network Address Translation
ISP	Internet Service Provider
API	Application Programming Interface
IIFE	Immediately Invoked Function Expression
TDD	Test-Driven Development

2. Introduction

Websites are strategic communication channels for customer relationships management, online marketing, and an interface to engage with the audience directly. Application areas are e-commerce platforms, help service providers, or information centers for knowledge, information, and products.

Many websites are implementing a web analytics solution to gather visitor-website interactions [Cli12]. These website interactions offer data for the optimization of the website to match visitor expectations [Ogl10; HMP09; CDN11]. Web analytics tools provide an information source for website stakeholders (for example, website designers, web developers, marketers, managers, e.g.) to adapt the website for visitor needs and align the website with the company's business goals [Kau10; Jan09; Ogl10].

Google Analytics is a prominent web analytics product to analyze the clickstream data produced by website visitors. Web analytics tools offer a holistic framework for the analysis of the clickstream data to reconstruct the user journey and identify performance indicators (PIs). The extraction of actionable insights is an essential research topic in web analytics, and many researchers are involved in the discovery of these insights [CCS12; KRS02]. The monitoring of key performance indicators or performance indicators provides data to understand changes in user behavior [Jan09; Has12]. Almost every web metrics provides quantified data that needs interpretation to discover information. In contrast, a contextual analysis of the search keywords offered by the internal search function provides interpretable and actionable information [Kau07; Has07; Jan09].

The Catrobat project offers teenagers a free and open-source development environment for the playful acquisition of development skills [Sla14]. The Catrobat project consists of development environments for the creation of applications and websites for the presentation of news and the sharing of created Pocket Code programs. The Pocket Code sharing website provides an environment for users and developers of the Pocket Code application to download and share their created programs. Furthermore, the website offers media resources, such as sample programs, tutorials, images, and sounds to support developers in the program creation. Features of the sharing websites are the modification, remixing, and re-uploading of created or downloaded programs. The sharing websites should offer an enjoyable environment for the program creation and the improvement of their development skills.

The Catrobat community website has implemented the standard page tag to monitor the website usage and to infer customer satisfaction. However, the data provided by Google Analytics was not used to interpret this collected data. This thesis introduces the concepts and web metrics to analyze the aggregated data and implements the event tracking feature to record on-site interaction with HTML elements and interactive content. Conclusive, the last chapter analyzes the collected clickstream data by page tags and event tracking with the previously introduced web metrics and customer journey. The last chapter gives a recommendation for the improvement of the Pocket Code sharing website based on the results of the web metrics and user journey. This recommendation suggests optimizations for the enhancement of the customer experience for the Catrobat community website's users.

Chapter 3 introduces the concept of business analytics and web analytics that are part of *Bussiness Intelligence & Analytics*. Web analytics processes a large volume of information. This volume of data is categorized through the umbrella term Big Data [GH15; FB13]. In recent times, this term has gained widespread popularity [GH15] and gets introduced in chapter 3. Furthermore, the chapter introduces the Trinity Strategy developed by Kaushik [Kau07]. The Trinity Strategy is a strategic approach to investigate a website's usage behavior through three different monitoring methods, the behavior analysis, the outcomes analysis, and the experience analysis.

Chapter 4 discusses methods to gather data about user interaction with the website. Three different data gathering methods are available, quantitative, qualitative, and competitive data collection methods [Kau10; BA07]. Web analytics tools are mainly using quantitative data for the analysis of website visits. Consequently, the chapter only outlines the qualitative and competitive data gathering methods. The most relevant quantitative data gathering methods for web analytics are the server-side log file analysis and the client-side page tagging method. Furthermore, the chapter discusses the advantages and disadvantages of these two data gathering methods. Lastly, the cookie technology is introduced and the differences between the individual types, as well as, their advantages and disadvantages are outlined.

Chapter 5 introduces the off-site and on-site metrics to gather quantitative clickstream data about website visitors. Data sources for off-site metrics are panel data studies, Internet service providers, and search engines. These off-site metrics gather the entire website traffic of Internet users and do not target a website principally [ZP15; KBR15; Cli15a]. In opposition, on-site metrics track user interaction of a specific website [Jan09; BBW+07; Bea13]. The on-site metrics can be categorized in four groups, the traffic source metrics (accessibility), the visitor metrics (characterization), the behavior metrics (navigation), and the content metrics (design/content) [Has07]. Chapter 6 introduces important web metrics, which are part of the four previously mentioned categories, as the data source for the website analysis.

The standard page tracking code is only able to gather information about the web page view [Cli12; Ana19c]. However, modern web applications are increasingly interactive and can update only partial content of a web page [NH05]. Such interactive events can be social media interactions, video interactions, reactions to events, outgoing links, or download of resources. Many websites and blogs are utilizing social media buttons to raise the engagement factor for the audience and to promote their websites [Has12; Kau10]. The page tag needs to be customized to track these previously mentioned interactions as introduced in chapter 7. Furthermore, chapter 7 outlines the differences between virtual pageviews and event tracking and discusses their commonly used application scenarios. Additionally, the chapter introduces the capturing of social media interaction and the calculation of the social engagement score.

Chapter 8 introduces the Catrobat project and the Pocket Code community website, as well as, their corresponding projects. These projects are the Pocket Code Android application, the HTML5 based development environment, Musicdroid, and Pocket Paint. Moreover, the chapter introduces the target group of the Catrobat project and illustrates the central web pages of the Pocket Code sharing website.

Chapter 9 introduces the trackable content of the Catrobat community website's web pages. These pages are the home page, the search page, the login page, the program page, the profile page, web pages of the media library, and web pages with embedded video content. The trackable content on the web pages is categorized in, page views, internal links, external links, and interactive content. Chapter 10 describes the current state of the clickstream tracking, and the possibilities of the user behavior analysis based on the gathered clickstream data. The chapter introduces the tracked content based on the four previously outlined categories of trackable content.

Chapter 11 discusses the structure of HTML documents and the code implementation of the tracking framework. One factor for a decreased page load speed is the repeated interruption of the web page rendering caused by the execution of JavaScript code. Chapter 11 introduces code optimization to increase the page load speed and the structure of the tracking framework. The tracking framework consists of the main module, the AnalyticsTracker, and the YoutubeTracker. These modules track page views, events of custom elements, outgoing links, emails, and video content.

Chapter 12 analyzes the clickstream data gathered by Google Analytics. The customer journey gets used to structure the user behavior analysis in three stages, the previsit-, visit- and postvisit-stage. Based on these three stages the visitor journey is analyzed. The previsit stage identifies the traffic sources. The visit-stage characterizes the website visitors and investigates their navigation path and viewed content. However, Google Analytics cannot gather information about the postvisit-stage. Consequently, no data for this stage is available.

The last chapter 13.1 summarizes the findings of the previous chapters and discusses the user behavior of the Pocket Code sharing website visitors. Furthermore, based on the findings of the web metrics, a recommendation is given for the optimization of the Pocket Code website, and further research topics are introduced.

3. Web Analytics

The "Internet", a network of global connected computing devices [DM13], is nowadays seen as a strategic business channel to increase revenue and attract visitors [McF05]. The Internet evolved from the in the late 1960s developed *ARPANET* (Advanced Research Project Agency Network) [LL05; Cas01] to a global computation network with more than 15 billion nodes in the year 2015 and is further expanding [Wee13]. The published paper in the year 1964 by *Lawrence G. Roberts* introduced the concept of packet switching to allow the communication of two computers via a computer network [Lei+09]. However, 25 years more were needed until Tim Berners-Lee published his thesis *Information Management: A Proposal* of a distributed hypertext system [Ber89; ANF12]. This proposal of a read-only web was the basis for the now available web technologies [ANF12]. Since then many improvements were made. The Internet revolutionized the exchange of information through the introduction of file sharing, social networking, blogs, instant message services and the World Wide Web (WWW) [Coh13]. The next step in the development of the web was the introduction of Web 2.0. Web 2.0 provides web users with the ability to participate in the exchange of information - the read-write web [ANF12]. As a result, the content of web pages did also evolve from static web pages to web pages with dynamically loaded content, so called *Rich Internet Applications* (RIAs) [NH05], such as Google Maps[1] or Bing Maps[2].

Rich Internet Applications are highly interactive and capable of nearly the same features as desktop applications [NH05]. The HTML5 web standard with the additional of web hooks enriches web application with multimedia content and instant message services [BI14]. However, compared to traditional web pages, Rich Internet Applications are not performing a complete

[1] `www.google.at/map`

[2] `www.bing.com/maps`

page load to update the content of a web page [Jan09; NH05]. Subsequently, with the evolution of the web, the amount of data gather via web analytics increased [McF05]. Data sources for web analytics tools are client-server interactions, web content, sentiments, surveys, geographic information, demographic information, search engines, or embedded sensors of mobile devices [CCS12].

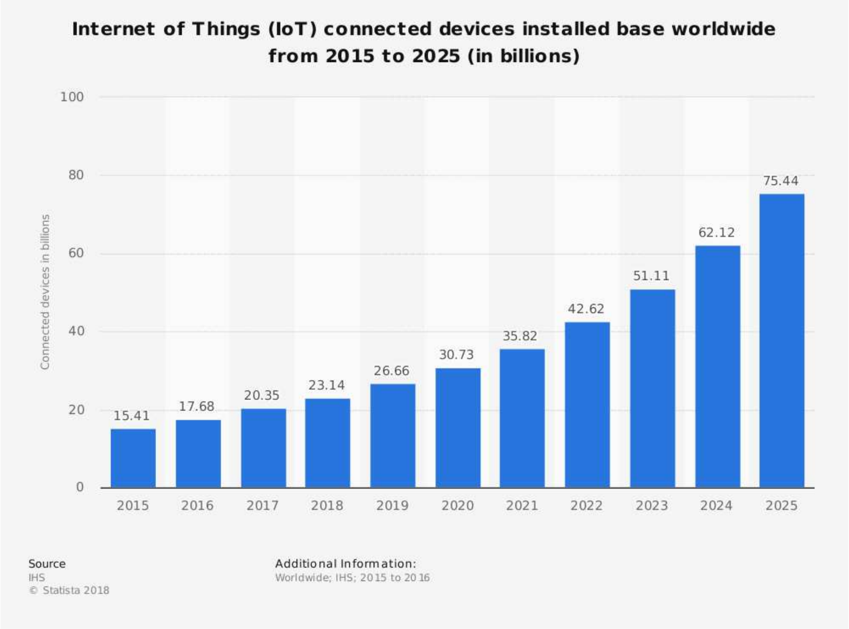


Figure 3.1.: Internet of Things (IoT) connected devices worldwide from 2015 to 2025 (in billions). Adapted from [Sta18]

Currently, 4.0 billion users[3] have access to the Internet. This amounts for 50% of the world population[4]. In the year 2010 only 1.9 billion people were able to access the Internet and 125 million websites were in use [Pan+11]. Figure 3.1 illustrates a forecast of the increasing amount of connected devices and shows the increasing trend of Internet connected devices. Market research from *Strategy Analytics* resulted in 33 billion Internet devices will be in use in the year 2020 [Mer14]. As Pani et al. [Pan+11] mentioned, the number of

[3] `http://www.internetlivestats.com/internet-users` Retrieved: 16 December 2018

[4] Total world population: approximately 7.6 billion people worldwide `http://www.worldometers.info/world-population` Retrieved: 16 December 2018

websites and visitors grows exponentially and the produced data correlates with the number of devices in use [FB13]. As a result, the scale of the processed data will also increase exponentially. This amount of data is classified under the term "Big Data" [GH15], which means a vast amount of data is available for the analysis in web analytics tools [CCS12; Hud+18].

3.1. Big Data

> *"Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis."*
> *(Gantz and Reinsel [GR11])*

The term *Big Data* was first mentioned by Silicon Graphics Inc. (SGI) in the mid-1990s [GH15; FB13]. However, only in the year 2011 did the term *Big Data* gain popularity, because global companies, such as IBM[5] or Google[6] started to invest in the analytics market [GH15]. The analyst Doug Laney significantly shaped the term *Big Data* as the combination of the three Vs: *Volume*, *Velocity*, and *Variety* [Lan01].

Volume The volume describes the size of the gathered data [GH15]. The boundaries of the volume are not clearly defined [GH15; FB13] and vary depending on time and application [GH15]. E-commerce companies and social networks are processing petabytes of data [MB+12]. For example, the size of the transaction history database of the retail company Walmart is estimated bigger than 40 petabytes and produces every hour 2.5 petabytes of data [Mar17; MB+12; Cuk10]. Social media networks are managing an enormous amount of information. In the year 2016, Facebook generated 4.6 billion "Likes" and 350 million photos are uploaded every day, resulting in approximately 126.8 billion photos every year [Low].

[5] www.ibm.com/

[6] www.google.com/

Variety The variety describes the type of gathered data [FB13]. The web content is distinguished in the *structured, semi-structured, and unstructured* content [GH15]. For example, social networks are producing data in the form of connection graphs, messages, likes, photos, and the direct interaction with the user interface [CCS12]. Emerging research of new data management technologies and analytics [GH15; FB13] incorporates data from mobile devices and embedded sensors [MB+12; FB13] to reveal new insights [Cuk10]. Clickstream data gathered from user interaction serves the purpose to provide web analysts information for the optimization of the website design [GH15] and to extract usage patterns [PC10].

Velocity The velocity describes the speed of the processing and reporting of analyzed data [KWG13; Kai+13]. Real-time processing and reporting of data offers an competitive advantage, because of the timely reaction to actual events [MB+12]. Subsequently, efficient system capabilities and algorithms are needed [FB13; KWG13] to process the input stream in real-time [GH15].

The three V's of Laney [Lan01] were further extended by three additional V's the *Veracity*, *Variability/Complexity*, and *Value*. The veracity indicates the imprecise and uncertain nature of data sources (for example, customer sentiments) [GH15]. The flow rate of the input stream is subjected to fluctuations, and the variability characterizes these changes [KWG13; GH15]. The complexity expresses the interconnectedness [Kai+13] and the challenge to connect, match, cleanse and transform the data from different sources [KWG13; GH15]. The last V, the value weights the information quality of the data source [Kai+13] in relation to its volume [GH15].

3.2. Analytics

This section is a brief summary of the history of analytics systems. Analytics frameworks, such as web analytics, are systems to *"analyze critical business data to help an enterprise better understand its business and market and make timely business decisions" (Chen, Chiang, and Storey [CCS12])*. The *IBM Tech Trends Report* of 2011 identified business analytics as one of the rising trends in information technologies [IBM11]. With the introduction of information systems in businesses, the capability to store, process, analyze, and display the collected information was featured [CCS12; CDN11]. The first business intelligence system, also known as *Business Intelligence & Analytics 1.0* (BI&A 1.0), did focus on data management of structured data often stored in relational database management systems [LCC13] with mostly company internal generated data [CCS12]. The concepts of BI&A 1.0 were introduced to convert raw data into relevant, usable and strategic knowledge and intelligence [Pir07] to support business decision making, the data-driven decision making (DDD) [Neg04; Ran09].

With the rising popularity and the increasing spread of HTTP-based web technologies [NZS06], such as the web and web services, new data sources of unstructured or semi-structured data became available [BA03; Neg04]. This data did not originate from in-house hosted data sources, but from user generated content [CCS12]. As a consequence, new research fields for data mining, such as text analytics, audio analytics, video analytics, social media analytics [GH15], network analytics, mobile analytics and web analytics [CCS12] emerged. However, only the topic of web analytics is further discussed in this thesis.

3.3. Web Analytics

With the publishing of the first web page, the question for website providers was *"How many visitors are viewing my content?" (Hassler [Has12])*. As a consequence, Dr. Stephen Turner wrote 1995 the first log file analyzer *Analog* [Kau07; Wu+09]. At the early stages, this analysis program had only limited analysis capabilities. By analyzing the web page requests recorded in log files, the *Hits* metric was generated [Has12]. At the beginning of the Internet, as a web page only consisted of text, this method of measurement was sufficient. However with the evolution of the web, observations based on information from *Hits* were insufficient to understand the user behavior [Kau07; Has12]. Nowadays a web page consists of multiple resources, such as JavaScript files, images, or CSS files [NH05] and the Hits metric counts every request of these resources. Due to the Hit metric inability to distinguish the type of the requested resource, a new metric the *Page View* was introduced [Has12; Kau07].

The Web Analytics Association[7] (WAA) proposed web analytics as:

> *"Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage"*
> *(Burby, Brown, Web Analytics Association (WAA) Standards Committee, et al. [BBW+07])*

Web analysis takes advantage of already existing methods of data mining and statistical analysis [CCS12] to understand visitor experiences [HMP09] and website interactions [CDN11; Has12]. Web analytics helps to understand website usage, but it does not offer tools to adapt the website for visitor needs [Wu+09]. Modern web analytics tools capture data about the visitor by observing the clickstream, the data input, or mouse movements [Kau07; Has12]. The gathered data is stored to discover usage patterns by applying web mining techniques to extract information [Pan+11; Fay+96]. Data mined information are the visitor path, the website usage [Wu+09], the completion rate of tasks [Ogl10], or conversion rates [Kau07]. On the one hand, the extracted information is used for the improvement of the website design [LCC13] by identifying choke points [Wu+09], which subsequently improves the performance of the website [PF13]. On the other hand, it is also used for marketing, identification of purchasing patterns, product placement optimization, and product recommendation [LCC13]. In summary, Web Analytics is used to understand user behavior and achieve an increase of customer satisfaction [NC11] by aligning web metrics with business goals [McF05].

Kaushik [Kau10] and *Burby and Atchison [BA07]* identified three categories of data sources: *quantitative*, *qualitative*, and *competitive* for web analytics. Quantitative data has the lowest information value and is collected by tracking the clickstream of website visitors [Kau07]. The clickstream describes the sequence of page views and interactions with the tracked website (for example, clicks, searches, or customized events) [Kau07; KBR15; Mon+04]. Web analytics tools are using this clickstream to extract information for

[7] `https://www.digitalanalyticsassociation.org`

Visits/Sessions, Visitors, Unique Visitors, Returning Visitors, Page Views, Landing Pages, Time on Site, Time on Page, Bounce Rate, Exit Rate, Exit Page, Traffic Source, or Events web metrics [Kau10; BBW+07; Sul+14]. Data sources for qualitative data are *Surveys*, *Voice of Customer*, *A/B Testing Lab Usability Tests*, or *Heuristic Evaluations* [Kau07; Has12]. Competitive data sources are *Panel data*, *Internet Service Provider*, or *Search Engines* [Kau07]. Information gained from quantitative data answers the "What?" questions of website usage, in opposition qualitative data the "Why?" questions [Kau07; BA07]. Due to the variety of information sources and the amount of available data, the challenge of web analytics is the extraction of *actionable insights* [LCC13; Kau10]. The aforementioned qualitative data investigates causes for a particular user-behavior extracted from the clickstream data. This includes information of performed clicks, the abandonment of the website, or the break-off of a started funnel [Kau07; Kau10]. Information obtained by web analytics is used to improve the design of the website [Ogl10; McF05], implementing new website features adapted for user needs, plan and adjust marketing strategies, asses the engagement value of the website content, or to reveal insight about user expectations [Ogl10; Kau10]. Optimization of the user experience increases customer satisfaction [PPC12] and improves website usability [Ogl10; WH06].

3.3.1. Trinity Strategy

Figure 3.2 illustrates the "*Trinity Strategy*" introduced by Kaushik [Kau07]. The Trinity Strategy is a strategic approach to extract *actionable insights* and define *observable goals* from collected information. It consists of the interconnected disciplines of the *Behavior Analysis*, the *Outcomes Analysis*, and the *Experience Analysis*. Each of these analyzes covers a subset of a user's website experience.

The *behavior analysis* analyzes a user's clickstream and reports information through *on-site metrics*. On-site metrics, which are discussed in chapter 5.3, are the foundation for the clickstream analysis. The goal of the behavior analysis is to *infer the intent* of a website visitor [Kau07]. Furthermore, the on-site metrics are collecting behavior, geographic, demographic, and technical information about every website visitor. This information serves the purpose

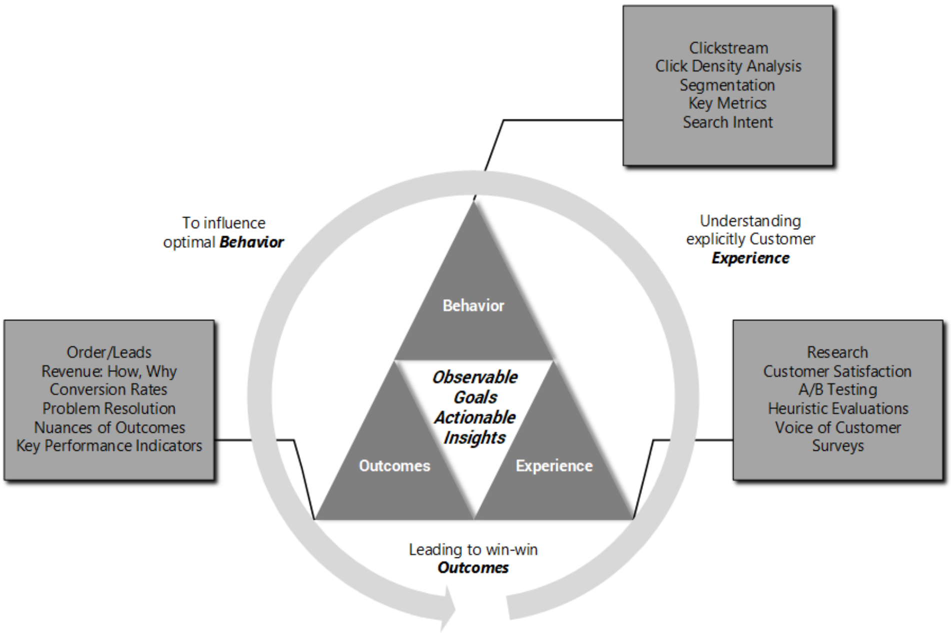


Figure 3.2.: Trinity Strategy. Adapted from [Kau07]

to segment the audience on multiple levels. For each segmentation level, the behavior analysis mines specialized insights of a user's behavior [Kau07]. By observing the search keywords, before and after a website visit, the behavior analysis is able to recreate a holistic user journey of website visitors [Kau10; Jan09]. In addition, the monitoring of internal search keywords provides insight into a visitor's behavior [Kau07] and helps to understand visitor needs [McF05].

The *outcomes analysis* measures the success rate and changes in organizational goals [Ogl10] by comparing the gathered information over time [Kau07; Jan09]. Outcomes are clear defined quantifiable indicators aligned with organizational goals [McF05; Jan09]. These outcomes are company dependent [Kau07; Has12]. For example, the success of e-commerce websites depends on revenue and sold products as such the outcomes are business oriented. Usually, non e-commerce websites are focusing on the frequency of visitors to measure the success of a website [Kau10; Lof12]. According to Kaushik [Kau10], the performance of non e-commerce websites is measured by the *Visitor Loyalty*, the *Visitor Recency*, the *Length of Visit*, and the *Depth of*

Visit metrics. Common outcomes used are *Conversion Rates, Key Performance Indicators (KPIs), or Click-through-Rates (CTR)* [Kau07; BS13]. The conversion rate is the percentage of visitors that are reaching a goal [Mey08]. KPI's are quantifying aspect of the user behavior to align the website goals with business goals [Jan09; Has12; McF05]. For an effective implementation of KPI's, they are required to be *S.M.A.R.T. Goals* [Has12]. S.M.A.R.T. Goals are defined as *"Specific, Measurable, Attainable, Relevant, and Timely" (Lawlor [Law12])* and used to identify areas of improvement, promoting web features, reviewing newly implemented functionality, and increasing revenue [Has12].

The behavior and outcome analysis are investigating the "What?" and "So what?" questions but for a holistic website analysis the "Why?" questions are needed to be answered, too [Kau07; Bea13]. The *experience analysis* complements the behavior and outcomes analysis by addressing these questions. It investigates motivations of user interaction and identifies areas of improvement [BA07]. *"Experience analysis allows us to get into the heads of our customers and gain insight or an a-ha about why they do the things they do" (Kaushik [Kau07]).* Furthermore, the experience analysis supports the testing of newly implemented features and designs in-house [Has12]. *Qualitative data* for the experience analysis is gathered by:

- ***Voice of Customer / Surveys***
  Usually, surveys are small questionnaires that enables the visitor to leave a feedback. The feedback is either an open-ended comment (Voice of Customer) or a Likert scale survey to comment the user experience of the current visit [Kau07].
- ***A/B Testing / Multivariate Tests***
  The A/B tested website delivers the visitor different variations of the same web page, a test version and a control version [Kau07; BA07]. The different versions are tested against a set of performance indicators, and the better performing variation will be further used [Has12]. Multivariate testing splits a web page in a number of modules, and testing tools deliver visitors with different combinations of these modules [Kau07; BA07]. Through statistical analysis, the best performing combination of modules will be determined [BA07].
- ***Lab Usability Testing***
  An impartial participant performs multiple tasks during a lab usability test. During this test, each comment, the problems a participant

encounters, the number of accomplished tasks, and the failures are captured. Lab usability tests are used to optimize the User Interface (UI) and workflows, but also to understand a customer's voice. [Tul+02; Kau07]

- ***Heuristic Evaluations***
  During a heuristic evaluation, researchers or web designers are performing multiple tasks. The evaluation of the website reveals every negative impact on the customer experience. The result of such an evaluation is used to optimize workflows, improve the UI, and to understand the overall level of usability. [CB10; Kau07]

The *Behavior Analysis*, the *Outcomes Analysis*, and the *Experience Analysis* are interconnected, and each covers a different aspect of a visitor's user experience to optimize the website for visitor needs. through iterative improvement of the website's usability [McF05]. All three analyses need to be considered to provide a website visitor the best user experience. Since every analysis is extensive, only the behavior analysis is further discussed in this thesis.

4. Data Collection Methods

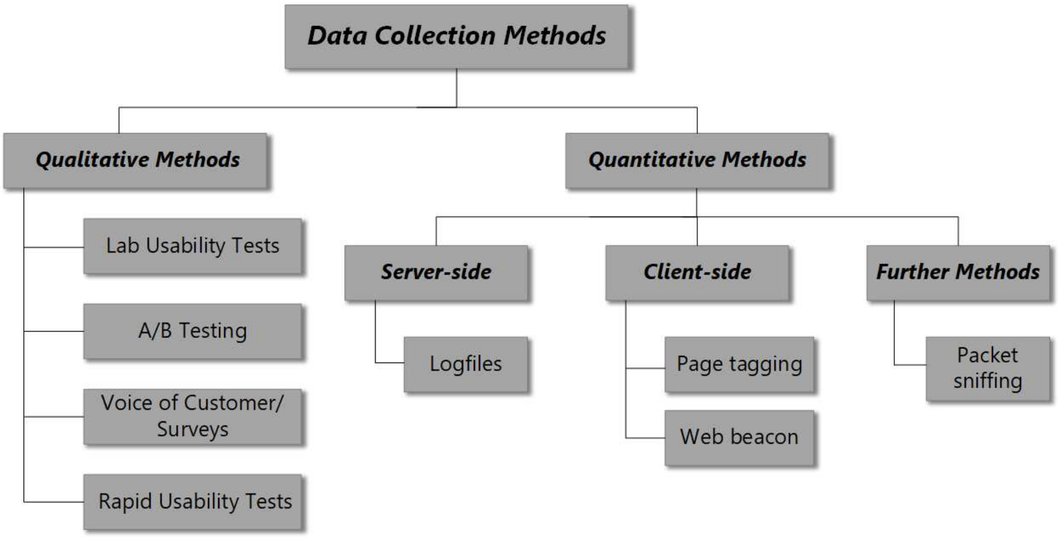


Figure 4.1.: Data Collection Methods. Adapted from [Kau07; Kau10; Has12; The09; BA07]

This chapter introduces the data gathering methods to track the clickstream data of website visitors. As in the previous section 3.3.1 discussed, the clickstream tracking is part of the behavior analysis. The clickstream provides data for the extraction of usage patterns [Pan+11] and to understand visitor experiences [HMP09]. The data collection methods are divided by their information value in: *qualitative data collection methods* and *quantitative data collection methods*, as shown in Figure 4.1. The quantitative data collection methods are differentiated in two major categories, *server-side* data collection methods and *client-side* data collection methods [HMP09; McF05]. Server-side data collection methods 4.1 are collecting data on a web server, and client-side data collection methods 4.2 on a visitor's device. In addition, the packet sniffing data collection method gathers data by intercepting the packet transmission of the client-server communication, as introduced in section 4.3.2.

The *server-side* data collection or *log file analysis*, records every transaction record from a server's perspective in a log file [Pan+11]. The first analytics tool *Analog* was written 1995 by Dr. Stephen Turner [Kau07] to extract the visitor's path through the website from a web server's log file [Pan+11]. However, the limitations of log files did predominate, and therefore web analytics tools did reduce the usage of log files as a data source [Kau07; Has12]. The *client-side* data collection methods are differentiated in *page tagging* and *web beacons*. Page tagging and web beacons are recording every action from a user's perspective. This chapter discusses the log file and page tagging data collection methods primarily. Further data collection methods, such as web beacons or packet sniffers are only outlined.

Qualitative data collection are *A/B Testing*, *Lab Usability Tests*, *Surveys*, *Rapid Usability Tests*, and *Heuristic Evaluations*. These qualitative methods are used to investigate the motivations of the audience to acquire information about their *needs, desires, and impressions* [Kau10; Has12; BA07]. In this thesis, the qualitative data collection methods are only briefly outlined in section 3.3.1 because the aim of this thesis is the investigation of techniques to analyze the clickstream data produced page tags of the Catrobat community website.

4.1. Server-side Log File based Data Collection

Log files are plain text files that are recording every transaction performed between a client and a web server [Has12; PC10; Bea13; SK09]. Log files contain every request from the client's browser to the web server [NC11; Pan+11]. However, web pages do not consist of a single resource, the web page, but also of embedded resources, such as images, JavaScript files, CSS files (Cascading Style Sheets) [Kau07; PC10; Has12; NH05]. The client performs for each embedded resource a request which is recorded by the log file [Kau07; PC10; Has12; NH05]. The log file even contains AJAX (Asynchronous JavaScript and XML) requests and downloaded files (for example, PDFs, images, or applications) [Bea13]. The process and the actors of the client-server communication and the lofile analysis are illustrated in Figure 4.2.

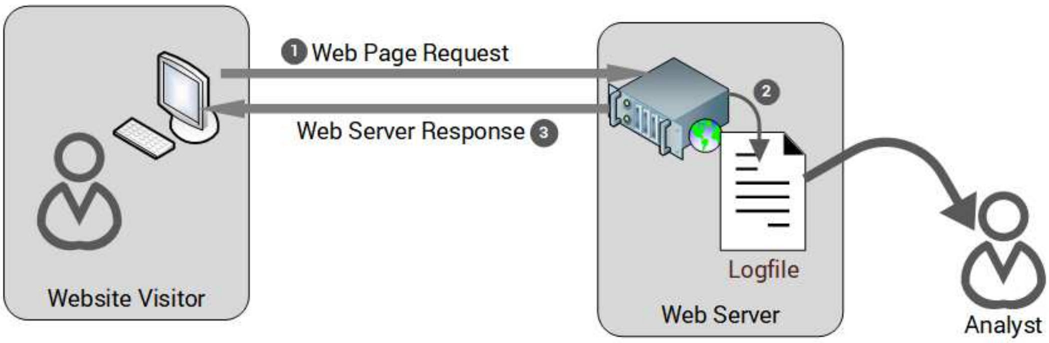


Figure 4.2.: Log file based Data Collection. Adapted from [Has12]

Actors *(illustrated in Figure 4.2)*
The actors of the log file data collection are the *website visitor/client*, the *web server*, and the *analyst*. A *website visitor/client* is a web browser, application (web API), or a web bot (spiders, robots or crawlers) that requests a resources from a web server. The *web server* receives these requests and returns the processed response. The *web analyst* needs direct access to the web server that stores the log files to conduct an analysis and is either an external organization or an internal employee specialized for the log files analysis [Kau07; Has12].

Procedure *(illustrated in Figure 4.2)*

1. Whenever a website visitor requests a resource/information from the web server or transmits a resource/information to the web server, the client sends a (GET, HEAD, POST, PUT, DELETE) request to the web server.
2. The web server records information about this request as an entry in the log file [Pan+11; SK09; KSK12]. This log file can either be stored locally or on a dedicated log file server [Jan09].
3. The web server's response includes the requested resource/information for the client.

4. Due to the log files data capturing architecture, the web analyst needs direct access to the log file server to analyze the information contained in the log files [Jan09; Kau07].

Modern web pages consist of various resource, such as the HTML file, images, JavaScript files, CSS files, or PDF files [Kau07; PC10; Has12; NH05]. Whenever an embedded resource of the web page gets requested, the web client creates a corresponding HTTP(S) request. As a result, the log file contains records for every embedded resource [Kau07; Pan+11]. These entries, essentially all embedded resources in a web page, are removed in the preprocessing phase of the log file analysis [Kau07; Pan+11; PC10; SK09]. Since these entries do not correspond to web page views, the data cleaning prevents the artificial increase of the page view metric [Kau07; Pan+11]. Furthermore, web bots, such as spiders, robots, or crawlers are independent operating programs [Hea06] that are requesting resources from a web server [Kau07; Pan+11]. For example, search engine bots are crawling the web for the generation of the search index of search engines [BP12; Bea13]. Traffic from such web bots can take up to 40% of the total amount of generated web traffic [Koh+04] and are identified and removed from the log file [Pan+11; SK09].

The log file technology was not introduced to collect data for the business decision making [Kau07] but to record web server errors [Kau07; Jan09]. However, due to the increasing importance to understand a website's user behavior, and the optimization of the website for the visitors' needs [Ogl10]. Web analysts did require detailed information about the user behavior of the website visitors [Jan09]. The *NCSA Common Log* log file format, illustrated in Table 4.1, gathers the IP address of the client, the client identifier, the date of the request, the HTTP method, the requested resource, the HTTP version, and the status code. However, the identification of subsequent web page request based on the NCSA Common Log has limitations (further discussed in section 4.1.2). As a result, the *NCSA Common Log*, illustrated in Table 4.1, was extended to the *NCSA Combined Log* format [Jan09]. Table 4.1 illustrates the *NCSA Combined Log* that records additionally the referrer, the user-agent, and the cookie. Web analysts can use this information for the analysis of visitor information, such as the navigation path, the web browser, or the operating system [Jan09]. The *NCSA Separate Log* distributes the content of the NCSA Combined Log in three separate files, the *Common log*, the *Referral*

Log, and the *Agent Log*. The *W3C Extended Log* extends the log files with user defined fields and identifiers [Pan+11; Jan09].

4.1.1. Advantages of Log File based Data Collection

The application of log files does not requires a specific web implementation or server configuration because a web server records web requests by default [PC10; Jan09]. Furthermore, the data aggregator (web server owner) possesses the ownership of the log file [Kau07]. This supports the fluent exchange of web analytics tools or vendors [Jan09]. Personal sensitive information contained in log files is normally kept confidential because the access to the log file server is restricted [Kau07]. Usually, only the web server owner or authorized personnel have access to the server [Kau07]. In opposition to JavaScript page tagging that embeds a JavaScript page tracking code in the web page, the log files based data gathering does not need a modification of the delivered web page to enable the data gathering [Jan09; Bea13]. Furthermore, log files are capturing activities of clients that do not execute JavaScript [Cli12; Bea13], such as web bots and Internet users with disabled JavaScript [Pan+11; Kau07].

4.1.2. Disadvantages of Log File based Data Collection

Log files were not introduced to gather data for business decisions [Kau07], but to log technical information, such as web server errors, the performance of the web server, or the web browser [Jan09]. As a consequence, extensive data processing is required for the removal of duplicate log file entries and the identification of cohesive web page views [Pan+11]. Many devices are using Network Address Translation (NAT) for the connection with the Internet [GFH14]. NAT devices are replacing the IP-addresses of multiple users with one public IP-address [Fom10]. Moreover, Internet Service Providers (ISPs) assign dynamic IP-addresses to customers [Has12]. As a consequence of NAT and dynamic IP-addresses, the log file analysis identifies multiple website visitors as a unique visitor [Has12; WH06]. Therefore, a new identification method was needed and found in the web cookies technology. The

Table 4.1.: Commonly used Log File Formats. Adapted from [Jan09]

Type of Log	Example Entry
NCSA Common Log	111.222.125.125 - jimjansen [10/Oct/2009:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043
NCSA Combined Log	111.222.125.125 - jimjansen [10/Oct/2009:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043 "http://ist.psu.edu/faculty_pages/jjansen/" "Mozilla/4.05 [en] (WinNT; I)" "USERID=CustomerA; IMPID=01234"
NCSA Separate Log	Common Log: 111.222.125.125 - jimjansen [10/Oct/2009:21:15:05 +0500] "GET/index.html HTTP/1.0" 200 1043 Referral Log: [10/Oct/2009:21:15:05 +0500] "http://ist.psu.edu/faculty_pages/jjansen/" Agent Log: [10/Oct/2009:21:15:05 +0500] "Microsoft Internet Explorer - 7.0"
W3C Extended Log	#Software: Microsoft Internet Information Services 6.0 #Version: 1.0 #Date: 2009 -05-24 20:18:01 #Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stemcs-uri-query sc-status sc-bytes cs-bytes time-taken cs(User-Agent)cs(Referrer) 2009-05-24 20:18:01 172.224.24.114 - 206.73.118.24 80 GET /Default.htm - 200 7930 248 31 Mozilla/4.0+ (compatible;+MSIE+7.01;+Windows+2000+Server) http://54.114.24.224/

recording of cookie information in the log file did improve the identification rate of returning visitors, as well as, the identification of continuous web page request [WH06; Has12; Cli12]. Since by default the web server only gathers the IP-address and the user-agent, the web server configuration needs to be adjusted to capture cookie information [Kau10; Jan09]. Furthermore, web browsers, firewalls, or web plugins can block or delete cookies, which decreases the accuracy of the unique visitor metric [Has12; AML07; Bea13].

The calculation of the *time on site* metric (discussed in section 6.6) requires the timestamp of the last viewed web page, as illustrated in Figure 6.6. However, the last web page request is an outgoing request and cannot be logged by log files, because the request is executed on the referring web server [WH06]. An incorrect calculation of the *time on site* metric is the result [WH06; BBW+07]. Further sources of log files inaccuracy are caching techniques, either by client caching, server caching, or proxy server caching [SK09; PC10; Mon+04]. Instead of performing a server operation that can be logged, caching techniques serve the requested resource from the cache [WH06; SK09]. As a consequence, the web server does not process the web request, and therefore the log file does not contain the entry of the request [Kau07; WH06; Bea13].

Nowadays, web applications are highly interactive and are including third-party content, such as video content [NH05; Bea13]. However, log files are not able to capture on-site interaction without a web request or interactions with third-party content. As a consequence, valuable information about visitor interactions is lost [Jan09; Bea13]. Nevertheless, websites that are highly using AJAX to update the website content, log files track these web request by default [Kau10; Bea13].

4.2. Client-side JavaScript Tag based Data Collection

The client-side JavaScript based data collection, page tagging, is the most popular method to collect data about website visitors [Cli12; Jan09; Bea13].

In opposition to the server-side data collection, which collects data from a server's viewpoint, the client-side data collection gathers data from a visitor's perspective [HMP09; Kau07] and supports a more detailed analysis of the on-site behavior of a visitor [Has12; Jan09].

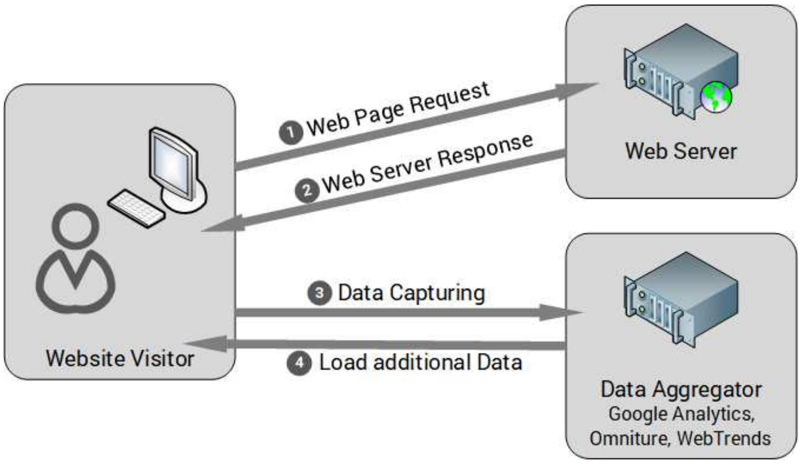


Figure 4.3.: Page Tag based Data Collection. Adapted from [Kau07]

Actors *(illustrated in Figure 4.3)*
The actors of the page tag data collection are the *website visitor/client*, the *web server*, and the *data aggregator*. Usually, a *website visitor/client* is a program that sends web requests to a web server and supports the execution of JavaScript, such as a web browser. Although, applications (web APIs), or web bots (spiders, robots or crawlers) are also able to request resources from a web server, page tagging is not able to gather data about these clients because they do not execute JavaScript [Cli12; Bea13; Pan+11], which is further discussed in section 4.2.2. The *web server* receives these requests from the visitor and returns the processed response. The *data aggregator* is either the website provider or a third-party web analytics vendor. Page tagging web analytics vendors, such as Google Analytics[8], Webtrends[9], or Adobe Analytics[10] are offered as SaaS (Software as a Service) for the capturing,

[8] `analytics.google.com`

[9] `www.webtrends.com`

[10] `www.adobe.com/data-analytics-cloud.html`

storing, and analyzing of the gathered data [Jan09; Bea13].

Web analytics tools are offered as *SaaS (Software as a Service)* [KSK12] and delivered by web services for an effortless integration in the existing IT infrastructure [GM09]. This supports the outsourcing of the analytics solution and the decoupling of the management department from the IT department [Kau07]. This decoupling simplifies the analytics process for small companies because they do not need to provide the necessary IT infrastructure to process the gathered data [DW07; KSK12]. Nevertheless, for privacy-sensitive data, the analytics solution should be hosted in-house [Kau07].

Procedure *(illustrated in Figure 4.3)*

1. Whenever a website visitor requests a web page from the web server, the client sends a request (GET, HEAD) to the web server.
2. The web server processes this request and returns the requested web page, identified by the URL. The response of the web server contains the web page with the embedded JavaScript tracking code, the page tag [Kau07; Has12; Bea13].
3. The web browser processes the received web page. During the web page load, the embedded tracking code gets executed [Bea13; KSK12]. This tracking code sends information about the currently viewed page (page view) and the visitor to the data aggregator [HMP09; Jan09; Bea13].
4. Furthermore, the data aggregator is able to modify the client by including additional data in the response [Kau07]. Usually, it is used to set a cookie in the client's web browser to track visitors across multiple sessions [Kau07; Jan09; Bea13].

The client-side data collection supports the collection of more detailed and accurate information about a visitor, than server-side data collection [Kau07; Has12]. Data collected via the client-side page tagging are [Has12]:

- Clicks
- Cursor position and cursor movement
- Screen resolution
- Browser window size
- Browser plug-ins

```
/* Global site tag (gtag.js) - Google Analytics */
<script async src="https://www.googletagmanager.com/gtag/js?id=GA_TRACKING_ID"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());

  gtag('config', 'GA_TRACKING_ID');
</script>
```

Listing 4.1: Google Analytics async Page Tagging. [Ana19c]

- Type and language of the browser
- Device
- URL
- Page title
- Referrer

Placement of the page tag The JavaScript tracking code or page tag must be embedded on the delivered HTML page to enable the client-side data collection [Cli12; KSK12; Has12]. Listing 4.1 displays an example tracking code of the Google Analytics page tagging. Nowadays, there are two different methods to load JavaScript code, *synchronous* and *asynchronous* [Mar11]. The technical differences of these two methods are further discussed in sections 11.2. The placement of the tracking code before the closing `</body>` tag of the HTML page in synchronous loaded scripts increases the loading speed of the web page [Cli12; Bea13; Mar11]. However, this also leads to missed page views because of the late execution of the tracking script [Cli12; Bea13]. Research done by *TagMan.com* (now *ensighten*[11]) did yield 20% more reported traffic, when the page tag is placed in the head element of the HTML page [Cli12]. The asynchronously executed JavaScript code does not influence the load time of a web page, and therefore the tracking code should be placed in the `<head>` element of the HTML page [Cli12; Bea13; Mar11].

[11] www.ensighten.com

In opposition to the synchronous tracking code, the asynchronous tracking code improves the performance of the page load time by approximately 5% to 10% [Cli12; Sha12].

4.2.1. Advantages of JavaScript Tag based Data Collection

Due to the client-side tracking, it is possible to track the user behavior from the visitor's viewpoint [Kau07; HMP09]. This supports the gathering of more detailed information about a visitor than the log file analysis [Has12; Jan09]. Furthermore, the data aggregator of page tags collects, processes, and reports the data in near real-time [Jan09].

The web consists heavily of interactive content, which ranges from videos, interactive web page functionality, dynamic loaded content, or WebGL [NH05]. Such interactions are trackable via the highly customizable page tag [Cli12]. Event tracking, further described in chapter 7, supports the tracking of these rich media content to provide information about the visitor behavior for informed web analytics [Has12] and business decisions [Kau07]. JavaScript page tags are under a constant technology development, for example, the transition from synchronous to asynchronous page tags, which improved the load time of web pages significantly [Cli12; Sha12]. Furthermore, it is possible to track visitors across multiple domains through the use of third-party cookies [Kau07; Jan09; Bea13].

4.2.2. Disadvantages of JavaScript Tag based Data Collection

A seamless implementation of the page tag on each page is necessary for the analysis of the user behavior. In opposition to the automatic data collection via log files [Pan+11; SK09; KSK12], page tags are placed manually on each web page by the webmaster [Mar11; Cli12]. Even if the website content is generated via a CMS (Content Management System) that supports the automatic placement of page tags, page tags can still be missing [Cli12].

Research done by analysts at MAXAMINE[12], did discover for business websites that even if a CMS is used for the placement of page tags, 20% of page tags are still missing [Cli12]. Furthermore, the high customizability of page tags is error prone and can lead to errors during the execution of page tags [Cli12]. Errors and missing page tags lead to a loss of valuable information about page visits [Kau07; Jan09; Cli12]. Page tags are blocked by Firewalls or web browser plugins, such as *Ghostery*[13], *NO Google Analytic*[14], or *NO-Script*[15] and as a result, no data about the clickstream is gather from these visitors [Kau07; Jan09; Cli12]. Furthermore, no information is gathered from clients that do not execute JavaScript, such as web bots (spiders, robots, crawlers) or web browsers with disabled JavaScript [Kau07; Jan09; Cli12].

A survey from the *Yahoo Developer Network*, in the year 2010, resulted that 1.3% of the users globally have JavaScript disabled [Zak10]. Newer statistics from the year 2013 done by *GDS*[16] (Government Digital Service) of the United Kingdoms (UK), did state, that 1.1% of users have JavaScript disabled [Ser13]. Consequently, a decreasing trend of users with disabled JavaScript is observable, which supports the utilization of page tags.

```
// JavaScript disabled Fallback
<noscript>
    <img src="http://www.google-analytics.com/collect?v=1&t=pageview&tid=UA-XXXXX-Y&cid=1234&dl=http%3A%2F%2Fwww.example.com&dt=Example%20Page%20Title" />
</noscript>
```

Listing 4.2: Google Analytics Web Beacon Fallback Solution [Ana19c]

Some web analytics vendors that are gathering information about a visitor by page tags, are providing a fallback solution to track page views even by disabled JavaScript [Kau07; WK09]. For example, this can be done via web

[12] www.maxamine.com, now Accenture Marketing Science www.accenture.com

[13] https://www.ghostery.com/

[14] https://addons.mozilla.org/en-US/firefox/addon/no-google-analytics/

[15] https://addons.mozilla.org/de/firefox/addon/noscript/

[16] https://gds.blog.gov.uk/

beacons, which are further discussed in section 4.3.1. Listing 4.2 displays an example code for a fallback solution for Google Analytics. The default implementation of page tags to gather page views is uncomplicated and only needs a few lines of code, displayed in Listing 4.1. However, the capturing of events like outgoing links, redirects, and downloads, requires a highly modified tracking code [Cli12; Has12].

4.3. Further Tracking Methods

Web beacons and packet sniffing are further methods to gather information about website visitors. Usually, these methods are not used for the click-stream or visitor path tracking, because the page tagging is highly marketed by web analytics vendors [Kau07; Bea13].

4.3.1. Web Beacons

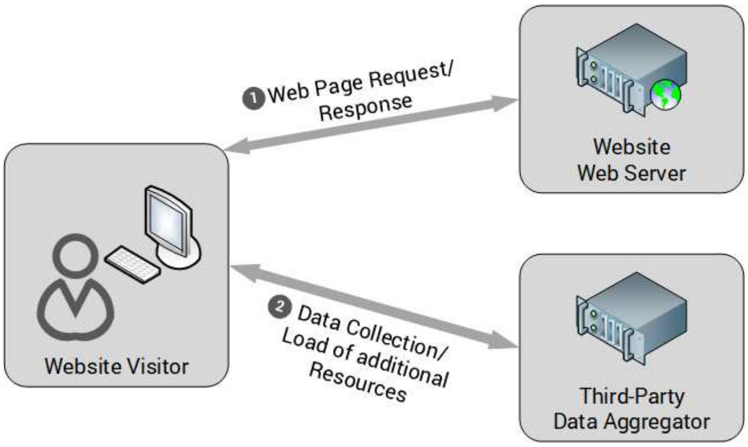

❶ Web Page Request/ Response
Website Web Server
Website Visitor
❷ Data Collection/ Load of additional Resources
Third-Party Data Aggregator

Figure 4.4.: Web Beacons Data Transaction Scheme. Adapted from [Kau07]

Web beacons are similar to JavaScript page tags and are gathering information about visitors on the client-side. Unlike JavaScript page tagging, that embeds a script tag in the HTML web page, web beacons are embedding a

HTML image tag in the HTML page to track website visitors, as illustrated in Listing 4.2 [Kau07; WK09; ZP15; SKS14]. This included image has the size of a transparent 1x1 pixel [WK09; SKS14]. Web beacons are often used by email campaigns and advertisement networks (for example, banner ads) to track users across multiple website domains [Kau07; WK09; KSK12; ZP15; SKS14]. As earlier mentioned, disabled JavaScript prevents the execution of the tracking code. In this case, some web analytics vendors are using web beacons as a fallback solution for the tracking of website visitors [Kau07; WK09].

Figure 4.4 displays the procedure of the data exchange for the tracking by web beacons.

1. Whenever a website visitor requests a web page from the web server, the client sends a request (GET, HEAD) to the web server. The website visitor sends a request of a web page, identified by the URL, to the web server. The web server processes this request and returns the requested web page, identified by the URL. The response of the web server contains the web page with the embedded image, the web beacon [Kau07; Has12]. However, the source attribute of this image does not target the same domain but a third-party domain [KSK12].
2. During the web page load, the embedded image gets requested from this third-party domain. This request sends information about the currently viewed page (the page view) and the visitor to the data aggregator [HMP09; Jan09; Bea13]. Furthermore, the data aggregator is able to modify the client by including additional data in the response [Kau07]. Usually, this is used to set a cookie in the client's web browser to track visitors across multiple sessions [Kau07; Jan09; KSK12].

4.3.2. Packet Sniffing

Packet sniffing is a rarely used method to gather data about website visitors because of the higher investment costs compared to other data gathering methods (for example, page tagging) [Kau07]. A packet sniffer is either a dedicated hardware device or middleware data layer that intercepts the data

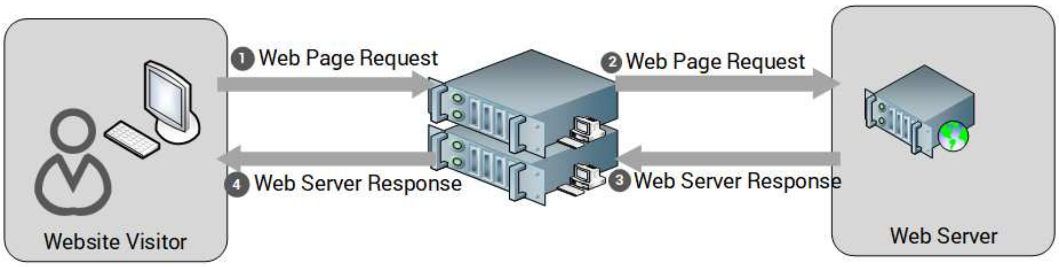


Figure 4.5.: Packet Sniffing Data Transaction Scheme. Adapted from [Kau07]

transmission between the client and the server [Kau07; KSK12]. However, only not encrypted network traffic can be analyzed. The missing network encryption causes privacy issues because packet sniffers are reading the entire content of the transmitted data via the client-server communication [Kau07]. This includes the content of the requests to the web server, the content of the responses from the web server, and cookie information [KSK12; Cus01]. Packet sniffing supports the analysis of the network traffic, for example, the response time of the web server, the content of web pages, requested resources, or send data to the web server [Kau07; Cus01]. Furthermore, packet sniffers are capable to modifying the sent content, for example, to embed page tags into every web page automatically [Kau07; WK09].

Figure 4.5 displays the procedure of the packet sniffing.

1. Whenever a website visitor requests a web page from the web server, the HTTP request is intercepted by the packet sniffer.
2. The packet sniffer analyzes the information contained in the HTTP request and forwards the request to the web server.
3. The web server processes this request and returns the processed HTTP response to the visitor. The packet sniffer intercepts this response and analyzes the content of the requested web page. Additionally, the packet sniffer can modify the content of the response and add the JavaScript page tag to a web page automatically. [Kau07; Cus01; KSK12; WK09]
4. Subsequently, the packet sniffer forwards the response to the client.

```
// Session Cookie
Set-Cookie: <cookie-name> = value
// Persistent Cookie
Set-Cookie: <cookie-name> = value; Expires = date
```

Listing 4.3: HTTP-Header Response to set Cookies. Source [W19]

4.4. HTTP Cookies

The HTTP-protocol is a stateless protocol [Fie+99], which means each request is independently processed and a web server cannot associate subsequent web requests to a single traffic source [Cli12]. Since the foundation for web analysis is the clickstream analysis [Kau10], techniques are needed to identify subsequent web requests of a unique visitor [Pan+11; KBR15; Sul+14]. HTTP-cookies, as defined in RFC6265[17] (HTTP State Management Mechanism), are small plain text files of name-value pairs [Cli12; KSK12] that are stored in the client's browser. The web server can use the information obtained from the cookie to match requests done by individual visitors [Kau07; Cli12; KSK12]. Page tagging tools are assigning a random generated *unique identifier* to each visitor. Web analytic tools are using this unique identifier to identify and track visitor requests in a single session (session cookie) and across multiple sessions (persistent cookie) [Kau07; Cli12; KSK12].

Cookies can be distinguished by the length of its validity in *session cookies* and *persistent cookies* [KSK12; W19]. Session cookies, also called *transient cookies*, are valid until the browser gets closed, which deletes the session cookie [Kau07; W19; KSK12]. The session cookie is used to identify the page views of an individual visitor during a single web session [KSK12; Sul+14]. Shown in Listing 4.3 is an example HTTP-header that sets a persistent and a session cookie. As displayed in Listing 4.3 the Expires or Max-Age attribute is not set in the HTTP header, which instructs the browser to delete the cookie during the browser termination [W19].

[17] https://tools.ietf.org/html/rfc6265

Persistent cookies do not get deleted after a session is terminated and are valid until the Expires date or the Max-Age is reached [W19]. Usually, persistent cookies have a deletion date far in the future [Kau07] to differentiate new visitors from returning visitors across multiple website visits [Kau07; KSK12; Cli12]. Web analytics tools are using the segmentation of the audience in new visitors and returning visitors to highlight the differences in the website usage of these two segments [PPC12]. The web metric of the *Unique Visitors* serves as information source in web reports for this segmentation, which is further discussed in section 6.2.

The identification of multiple visits via persistent cookies is influenced by various factors [BA07; Bea13]. For example, a study of comScore[18] did conclude that 31% of Internet users are deleting their first-party cookies monthly [AML07]. This is either done manually through browser options or automatically by Firewalls, AntiVirus programs, or by browser settings [AML07]. As a result, the returning visitors are identified as new visitors because of the missing persistent cookie.

Additionally, cookies are differentiated in *first-party* and *third-party* cookies [BA07; Kau07]. *First-party cookies* have the same domain name as the viewed website that is identified by the URL. A study done by *Stone Temple Consulting Corporation* did discover, that the deletion and rejection rate of first-party cookies is 13% lower than the deletion and rejection rate of third-party cookies [Eng07]. Thus it is recommended to use first-party cookies for the identification of website visitors [Cli12; Kau07; Eng07]. *Third-party cookies* have a different domain name in relation to the currently viewed website. Usually, third-party cookies are used to track users across multiple domains, for example, advertisement networks, social media networks, or video providers [MM12]. Furthermore, third-party cookies can be viewed as an privacy intrusion of users [Kau07; MM12]. Some AntiSpyware programs or Firewalls are detecting and blocking third-party cookies automatically, or the user deletes the third-party cookies manually [Cli12; Eng07].

[18] `http://www.comscore.com`

5. Web Analytics Metrics

In this chapter, the difference between *off-site metrics* and *on-site metrics* are discussed and the commonly used on-site web metrics are introduced. Off-site metrics are only briefly discussed because the focus of this thesis is on on-site metrics to analyze the clickstream behavior of website users. Discussed is how on-site metrics can be utilized to mine information about the audience and the key points of each metrics are outlined. The metrics are categorized in four categories, the traffic source metrics (accessibility), the visitor metrics (characterization), the behavior metrics (navigation), and the content metrics (design/content).

5.1. Off-site vs On-site Metrics

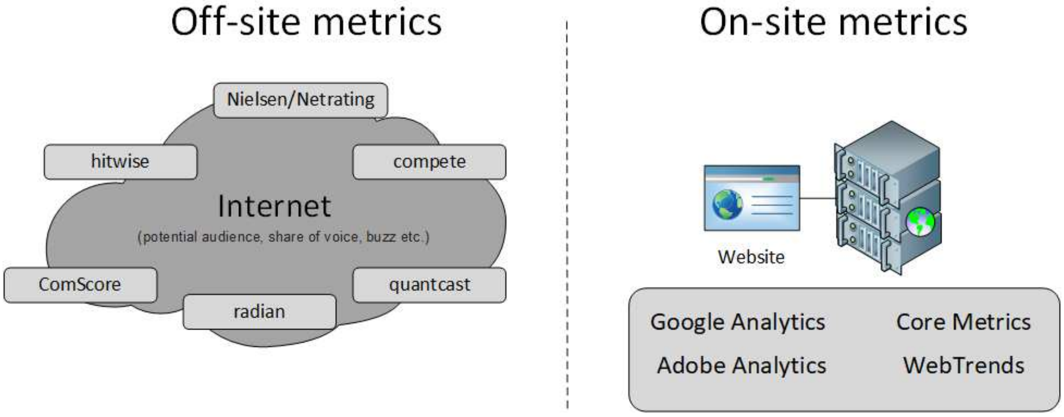


Figure 5.1.: Off-site and On-site Metrics. Adapted from [Cli15b]

Off-site and on-site metrics provide data about the clickstream of website visitors as illustrated in Figure 5.1. However, on-site metrics gather data via direct interactions with the tracked website (for example, from page tagging or log files) and off-site metrics collect the entire clickstream and browsing behavior of Internet users (for example, from panel data or ISP) [ZP15; KBR15; Cli15a]. The data capturing methodology and techniques are different for off-site and on-site metrics. Therefore, the values of different web metrics, such as the number of visitors or the number of page views are not comparable [Cli12]. However, meta information derived from web metrics such as visitor trends are correlating [Kau10].

5.2. Off-site Metrics

Off-site metrics are used to gather information about the entire browsing behavior of Internet users [Kau07; Cli12; KBR15]. This includes information remotely to the analyzed website but also data that is comprised of the analyzed website, for example, the entire path through the Internet, search keywords, or social media mentions [Cli12; KBR15]. Information mined from off-site metrics are the size of the *potential audience* (market potential), the *visibility of the website* (share of voice), or the *sentiment of the audience* (buzz) [Kau10; Cli12]. Additionally, information obtained about the website are *performance benchmarks*, an indication of the *effectiveness of search engine campaigns*, or insights into the *user behavior of visitors on competitor websites* [Kau07; Cli15c]. This competitive information serves the purpose to compare performance indicators of the tracked website with the competition [Kau10; Cli15a]. Furthermore, some off-site analytics vendors relate demographic information about the potential audience to the gathered data [Cli15b], which is discussed in the individual sections of the off-site information sources.

The data for off-site metrics can be acquired by:

- Panel Data Study
- Internet Service Provider (ISP)
- Search Engines
- Social Media Networks

5.2.1. Panel Data Study

Voluntary participants of a panel data study are installing a monitoring software on their devices [Kau10]. This monitoring software funnels the entire network traffic through proxy servers of the panel data company [Kau07]. The panel study company analyzed this gathered information to offer observations about the entire Internet browsing behavior of each study participant [Kau07; Hsi14]. This information is compromised of similar visited websites, the browsing path, search keywords, and interests of visitors [Kau07; KBR15]. However, since the monitoring software tracks the entire web traffic, the company also captures sensitive information like names, addresses, credit card numbers, or social security numbers [Kau10; MM12]. This prevents the usage of a panel data study in privacy-sensitive environments [Kau07].

The number of participants in a panel data study ranges from a few thousand to more than a hundred thousand participants [Cli15c]. Panel data study companies are mostly located in the United States (US), for example, ComScore[19], or Nielsen Netratings[20]. As a result, the accuracy of the provided data decreases for other geographic location [Kau07]. Since the user in a panel study is known, demographic information about the user, such as age, gender, ethnicity, or income bracket is available [Cli15c]. A panel data study does not target a specific website. As a result, off-site information is not provided for every website [Kau07; Cli15c].

5.2.2. Internet Service Provider

Internet service providers (ISPs) are capturing every user transaction record of each connected device [Kau10]. This information is then stored in server log files [Kau10]. The analytics vendor aggregates and analyzes the received data to provide insights into the potential audience, the visibility, or the sentiment [Kau10; Cli12]. The ISP anonymizes the collected data records before they are sold to analytics vendors to ensure a high level of privacy,

[19] http://www.comscore.com/

[20] http://www.nielsen.com

for example, Hitwise[21] [Cli15c; BHF12]. As a consequence, no demographic information associated with the collected data is available [Kau10; Cli15c]. The sample size of the gathered data is significantly larger (millions) and more diverse than the sample size from panel data (hundred thousands), which results in a higher accuracy of the provided information [Kau07; Cli15a]. However, off-site metrics are not able to provide an in-depth analysis of the user behavior of a website [Kau07]. Only on-site metrics analyze the entire clickstream of websites [Kau07; Jan09; Bea13].

5.2.3. Search Engines

Many users are using search engines [BB98] to navigate the Internet [Bro02] and search for topics of interest [Kau07]. Popular search engines are Google[22], Bing [23], Yahoo! Search[24], Ask[25], or DuckDuckGo[26] [eBi17]. Search engines are generating data in the form of search queries by tracking the input of the users. Data mining techniques can be used to reveal the visibility of specific websites, the identification of landing pages, or collective trends [RGS09]. Search trends, offered by services such as Google Trends[27], provide in-depth information about currently popular keywords of the potential audience [Kau07; RGS09]. Search engines are gathering the search keywords of a visitor, before a visit, to arrive at the website, and after a website visit. An analysis of these keywords reveals competitive information for strategic business and marketing decisions [Kau10; RGS09]. Furthermore, search engines are relating demographic information to search engine users either through the browsing behavior [BHF12] or by matching users to social network profiles [Kau07]. Based on this information, a demographic profile of the potential audience can be created [Kau07; RGS09].

[21] http://connexity.com/hitwise/

[22] https://www.google.com/

[23] https://www.bing.com/

[24] https://www.search.yahoo.com/

[25] https://www.ask.com/

[26] https://duckduckgo.com

[27] https://trends.google.com/trends/

5.3. On-site Metrics

On-site metrics measure every website activity on the website, such as the clickstream, the visitor path, visitor information, web funnels, the website performance, the website design, conversion rates, or website events [Cli12; HMP09; Jan09]. The first touchpoint with the website is usually referred as the *landing page* and the last touchpoint of interaction the *exit page* [Kau10; BBW+07; Bea13]. During the entire user journey, a website visitor produces data that is gathered and analyzed by web analytics tools to improve the website design, website navigation, website performance [HMP09], or to identify the consumed content [JK15]. Consequently, the on-site metrics measure the performance of a website [BA07] to adjust the website for the visitor needs [Ogl10; HMP09].

Furthermore, on-site metrics are able to identify the methods of a visitor's arrival, the traffic sources. Traffic sources are the direct traffic (bookmarks, direct input of the URL), the organic search traffic (search results), the referral traffic (references or links on third-party websites) [Cli15a], or advertisements hosted on third-party websites [Kau10]. On-site metrics are used as information source to measure the success of KPIs in regards to a company's business goals [SKS14; Kau07; KSK12]. For example, the CTR (Click-through-Rate) is a performance indicator for advertisements [Kau10]. and measures the performance of marketing campaigns [RDR07]. Additionally, on-site metrics are tracking website events and user interactions with rich media content via customized tracking code, as discussed in chapter 7.

5.3.1. Clickstream Analysis Categories

Studies by Peacock [Pea02], Xue [Xue04], and Yeadon [Yea01] have identified web metrics to improve a website's content, the navigation, the design, and to evaluate a website's accessibility [HMP09]. Web metrics to improve the content of a website are the exit page [Pea02], search keywords (6.19), referrer (6.17), search engines (6.18), top entry (6.12) and exit pages (6.13), or the time on site metric (6.6) [Xue04]. The error pages, search keywords [Pea02] (6.10), and behavior flow [Pea02; Yea01] (6.9) are used to improve the navigation of

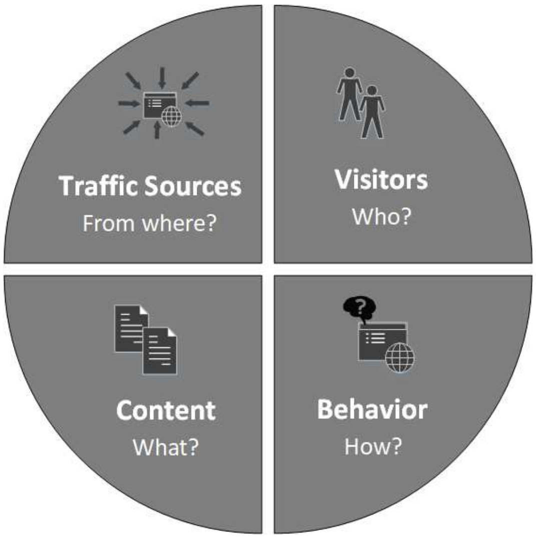


Figure 5.2.: Clickstream Analysis Categories by Marco Hassler. Adapted from [Has12]

a website. The browser [Pea02; Yea01] and device information (6.5) [Pea02] provide information to improve a website's design. Search keywords [Pea02; Xue04; Yea01] (6.19), search engines [Xue04] (6.18), referrer [Pea02; Yea01] (6.17), or landing pages [Pea02] (6.12) evaluate a website's accessibility.

A study by Pakkala, Presser, and Christensen [PPC12] highlights the importance of the user differentiation in new visitors (6.1) and returning visitors (6.2). Moreover, Hassler [Has12] provides a holistic framework for the clickstream analysis and integrates the aforementioned aspects of the content, the navigation, the design, and the accessibility. The framework differentiates the web metrics in four categories, the *visitor analysis*, the *behavior analysis*, the *content analysis*, and the *traffic source analysis*, as illustrated in Figure 5.2. Each of these categories monitors a different aspect of a website's visit, which are in the following section described. Furthermore, information mined from these web metrics can be used to segment the audience and to identify differences their user behavior [BG15; Kau07; Web13a].

Visitor Analysis The metrics and reports of the visitor analysis attempt to identify characteristics of the website visitors. The visitor analysis includes the elemental metrics of each web analysis, the *visits* metric and *visitors* metric [Kau07; Has12; BS13], which are further discussed in section 6.1 and section 6.2. Visitors are unique entities, such as different devices or browsers, that have one or multiple website visits [Kau10; BBW+07]. Visits or sessions designate a behavior of the visitors and consist of the following three website visit stages: the *landing stage*, the *browsing stage*, and the *exit stage*, which is further discussed in chapter 12. Metrics and reports of the visitor analysis are determining the frequency of the website visits per unique visitor, as outlined in section 6.3. Additionally, the audience of a website can be segmented by demographic traits or geographic characteristics, as described in section 6.4, as well as, the used technology to reach the website, further addressed in chapter 5.3.1.

Behavior Analysis Metrics and reports of the behavior analysis are identifying the user behavior and characteristics of a website's visit. This includes the duration of a visit (6.6), the number of web pages visited during a single web session (6.7), the number of visitors with no website interaction (bounce rate) (6.8), the visitor path through the website (6.9), and the usage of the internal search function (6.10). This information can be used to segment the website's audience based on their behavior to examine the differences of these segments [Has12; Web13a].

Content Analysis The content analysis identifies the consumed content of website visitors [Has12]. This includes an analysis of the consuming behavior, such as the most frequented web pages (6.11), key web pages, landing pages (6.12), and exit pages (6.13), which is further discussed in chapter 6. Furthermore, the content analysis analyzes the loading time of web pages (6.14), as well as, the consumed video content (6.15).

Traffic Sources Analysis The information gathered by metrics and reports of the traffic source analysis identifies the traffic sources and monitors their performance [Has12], as illustrated in Figure 5.3. The traffic channel types are:

- Direct traffic

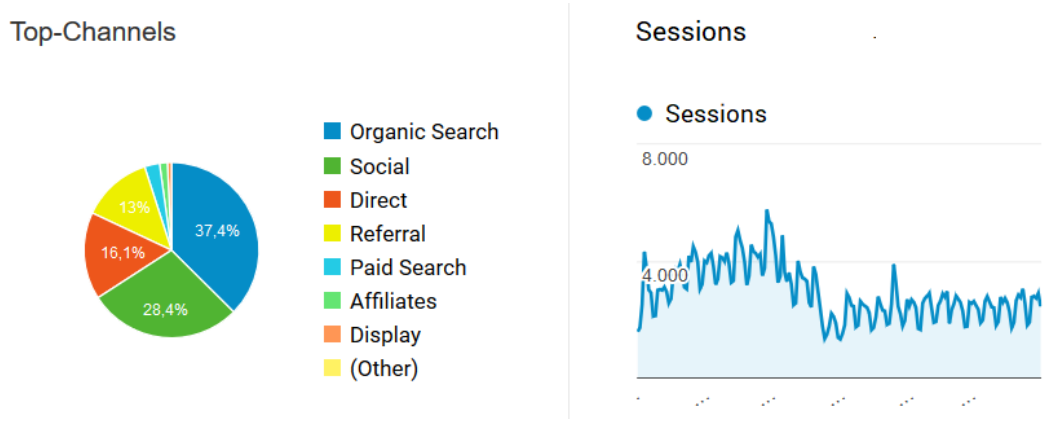


Figure 5.3.: Top-Channels from Google Merchandise Store. Source: Screenshot from Google Analytics of Google Merchandise Store (`www.googlemerchandisestore.com`)

- Referral traffic
- Search engines
  - Organic search
  - Paid search
- Social networks
- Paid advertisement
- Other types of advertisement (offline advertisements, et cetera)

The gathered information of the traffic source analysis is used to optimize these traffic source, and to adjust current or plan future marketing campaigns in correspondence with the performance of previous campaigns [Has12; JK15]. Furthermore, information gathered by the traffic source analysis can be used to segment the visitors based on the traffic source [Web13a; BG15]. In chapter 6, the aforementioned metrics and reports of the traffic sources analysis, such as the direct traffic metric (6.16), the referral traffic metric (6.17), the search engine traffic metric (6.18), the search engine keywords metric (6.19), and the social media traffic metric (6.20) are introduced.

6. On-site Metrics and Reports

In the following chapter, frequently used web metrics and reports that are gathering information about the audience are introduced. The discussed web metrics include the web metrics mentioned in section 5.3.1. At first, the chapter introduces the essential web metrics that are part of every web report, the visits metric and the unique visitors metric. Furthermore, since cookies can be blocked, deleted, or be unavailable, a data inaccuracy exists in the measurement of unique visitors. This chapter addresses this data inaccuracy, followed by web metrics that are reporting data about the visitor, the visit, geographic information, technical information, navigation behavior, consumed content, or traffic sources.

6.1. Visit

A visit or also called a session is a collection of subsequent web requests and begins with the arrival on the website and ends with departure from the website [Kau07; Bea13; Has12; BBW+07]. The arrival on a website is determined by the first interaction with the website. However, web analytics tools are not able to unmistakably determined a website's departure, because a session is automatically terminated if the user does not interact with the website for a certain amount of time [Bea13; Kau10]. If a user returns before the session is expired (usually 30 min), the session will get continued or else a new session is started [Bea13; Kau10]. The measurement of the session duration is further discussed in section 6.6. Web analytics tools are measuring a visit by generating a unique session-ID during the first web request to the web server. This session-ID is saved in a session cookie on the client's browser, discussed in section 4.4, and included in each following web requests to distinguish subsequent web page requests from one visitor during a visit [Ana19c].

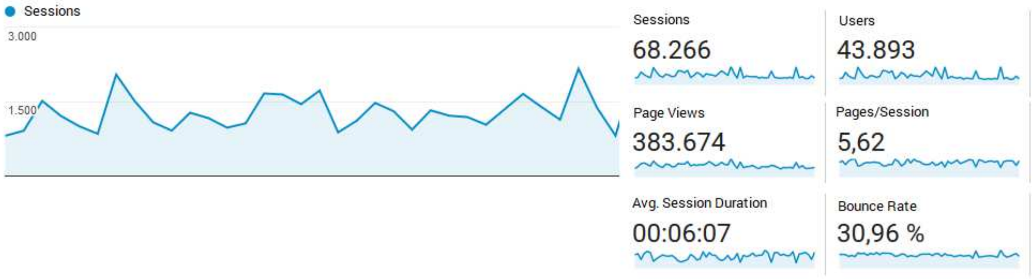


Figure 6.1.: Visits Metrics. Source Screenshot from Google Analytics of `share.catrob.at`

The visits metric displays the total amount of sessions, as illustrated in Figure 6.1. However, unaccompanied the visits metric provides only the number of visits, but in combination with segmentation, the visits metric supports the investigation of the performance of each individual segment [Bea13; Has12]. For example, a segmentation of the audience in new visitors and returning visitors, or different time periods can be compared to measure the number of visits [Has12].

An observation of the visits metric is able to identify performance changes in the number of visits. This serves the purpose to identify changes induced by *Search Engine Optimizations (SEO)* [Zil15], the impact of *Pay-per-Click Advertisements (PPC)*, the effects of social media marketing, or other advertisement campaigns [Bon12]. Since the number of visits depends on the observed time period, for example, the daytime, the week, or holidays, it is recommended to compare similar time periods or regions to reduce data inaccuracy [Has12].

6.2. Visitors - Unique Visitors

The unique visitor metric estimates the number of individual users that are visiting a website one or multiple times [Kau10; BBW+07; BS13]. Based on the information gained from the unique visitor metrics, the visitors are distinguishable in new visitors and returning visitors, as shown in Figure 6.2. Furthermore, Figure 6.2 displays the total number of sessions, as well as, the ratio of visits from new visitors to returning visitors.

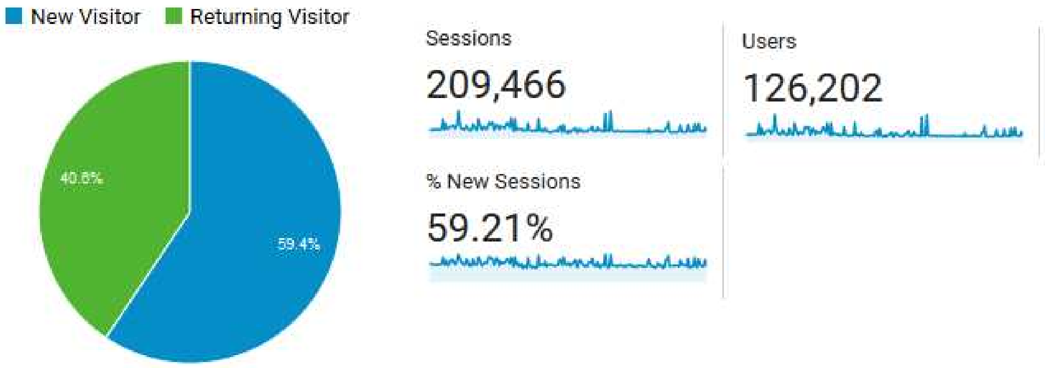


Figure 6.2.: Unique Visitors Metrics. Source: Screenshot from Google Analytics of `share.catrob.at`

The identification of unique visitors is similar to the identification of visits, but instead of setting a session cookie during the first page view, a *persistent cookie*, discussed in section 4.4, is set in the client's web browser [Ana19c]. This persistent cookie does not expire after the visitor closes the browser or the web session is terminated. It consists of a unique identifier to identify returning visitors [Ana18] Furthermore, the cookie does not contain any personal information and is solely used to identifier returning visitors [Kau10; Ana18].

The unique visitor metrics is used to determine the ratio between new visitors and returning visitors, and to segment the audience to analyze differences in the user behavior between these two segments [Has12; PPC12]. For example, the number of page views per visit, certain interactions with the website, or differences in exit pages [PPC12].

Data accuracy of unique visitors As in section 4.2 and section 4.4 outlined, page tagging with cookies cannot absolute identify returning visitors, which leads to data inaccuracy. The most common factors, that can impact the data accuracy of the unique visitor metrics are [Eng07; Cli12; Kau10]:

```
// Google Page Tag - Set User ID
gtag('config', 'GA_MEASUREMENT_ID', {
  'user_id': 'USER_ID'
});
```

Listing 6.1: Google Analytics User ID Tracking. Source: `https://developers.google.com/analytics/devguides/collection/gtagjs/cookies-user-id`

- Rejected or deleted cookies
- Same visitor on multiple devices
- Different users on the same device
- Same device with different browsers

Rejected or deleted cookies As in section 4.4 discussed, Internet users are able to install tracking-blockers to reject cookies. Since the rejection rate of third-party cookies, because of a firewall, browser settings, or installed plugins, is higher than the rejection rate of first-party cookies, it is recommended to use first-party cookies to track unique visitors [AML07; Has12; Kau07; Eng07]. Subsequently, the rejection or deletion reduces the accuracy of the unique visitor metric [Cli12; Kau07; Bea13]. A further source for the data inaccuracy is the web browsers' deletion of stored cookies or the expiration of cookies [Eng07; Cli12]. Furthermore, the privacy mode of web browsers also deletes persistently stored session data.

Same visitor on multiple devices A unique visitor is able to visit a website from multiple devices, for example, a smartphone, a tablet, a work-station, or a home-station. As a consequence, each device is identified as a different unique visitor because page tagging web analytics tools are tracking cookies, not visitors [Kau07; Ana19b; Bea13]. Web analytics tools are offering a method to add additional information to the page tag to identify returning visitors, for example, a login identifier [Cli12; Ana19b]. The Google Analytic API is able to identify a logged-in user by customizing the tracking code and adding a user id to the page tag. Listing 6.1 shows an example code for the tracking of additional information, such as the user id.

Different visitors on the same device Different people can share the same device, for example a family PC. This device-sharing leads to a false identification of different visitors as one unique visitor because the unique persistent cookie gets set per device, not per visitor [Cli12; Ana19b]. As previously mentioned, web analytics tools are tracking cookie, not visitors [Kau07; Ana19b; Bea13]. This leads to different identified user-behaviors of the same unique visitor [Cli12; Ana19b]. As a consequence, an additional identification method must be used to prevent this misidentification. For example, the tracking of logged-in users by adding a user id to the page tag [Ana19b]. However, if the tracked website does not support a user-management, or if users are using the same login session, the unique visitor cannot be identified.

Same computer different browsers used A visitor can use different browsers to visit the same website. Since each browser stores cookies individually and cookies are not shared across browsers [Aye+11], this leads to a false identification of a single unique visitor as two different unique visitors [Kau10; Cli12]. As a result, a unique visitor is identified as two different unique users. Like previously described, identification across multiple devices or browser needs an additional identification method, such as a login.

Each analytic tool uses a different approach to compensate for the inaccuracy in the unique visitor measurement, which leads to differences in the numbers [Kau10; Bea13; Has12]. This difference can range from 10% to 20% [Cli12; Bea13]. However, the segmentation in new visitors and returning visitors or the identification of visitor trends is more important than the numerical value of the unique visitors metric [BS13; Has12; Bea13].

Table 6.1.: Unique Visitors Calculation. Adapted from [Kau10]

	Visitors			Weekly Unique Visitors
Week 1	Visitor A	Visitor B	Visitor B	2
Week 2	Visitor B	Visitor B	Visitor C	2
Absolute number of Unique Visitors				3

The unique visitor metric provides information about the number of unique visitors in a certain time period [BBW+07; BG15]. However, this information is time dependent and the number of unique visitors from different time periods cannot be added [BBW+07; Kau10]. Table 6.1 demonstrates an example calculation of the number of unique visitors. Week 1 and week 2 have two unique visitors respectively, added together this results in four unique visitors. However, the total number of unique visitors is three (Visitor A, Visitor B, and Visitor C). Reports in web analytics tools are configurable to observe a certain time period to provide the number of unique visitors for any given time period.

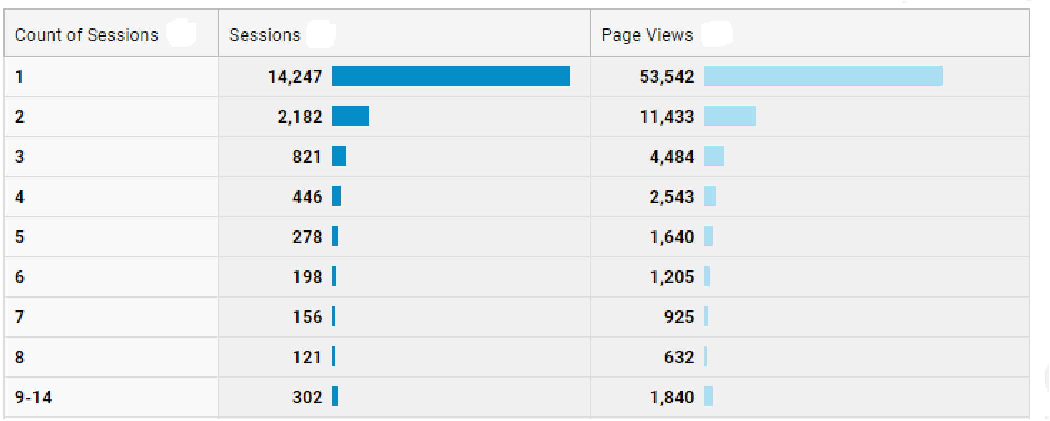
6.3. Visitor Loyalty

The visitor loyalty is an important indicator for the success of non-profit websites [Kau10]. It consists of two different reports the *Count of Sessions* report that is illustrated in Figure 6.3 and the *Days Since Last Visit* report. Both reports are gathering information about a unique visitor's user behavior [Lew13] and are providing insight into the satisfaction of user expectations [Ken+11].

The inaccuracy of the unique visitor metric, mentioned in section 6.2, subsequently decreases the accuracy of the visitor loyalty reports [Has12; Cli12]. This inaccuracy is cumulative and increases with the length of the observed time period [Has12]. As a consequence, the real amount of returning visitors is higher than the observed number in the reports [Cli12].

6.3.1. Count of Sessions

The *Count of Sessions* report observes the number of sessions done by a single visitor. The web metric is used to understand the visiting patterns of website visitors [Lew13] and to indicate the visitors loyalty by observing the number of sessions [Has12; Kau10]. Furthermore, the count of sessions metric identifies volatile changes in the usage behavior of the website [Kau10]. Visitors with only one session are occupying a big part of the diagram, as



Count of Sessions	Sessions	Page Views
1	14,247	53,542
2	2,182	11,433
3	821	4,484
4	446	2,543
5	278	1,640
6	198	1,205
7	156	925
8	121	632
9-14	302	1,840

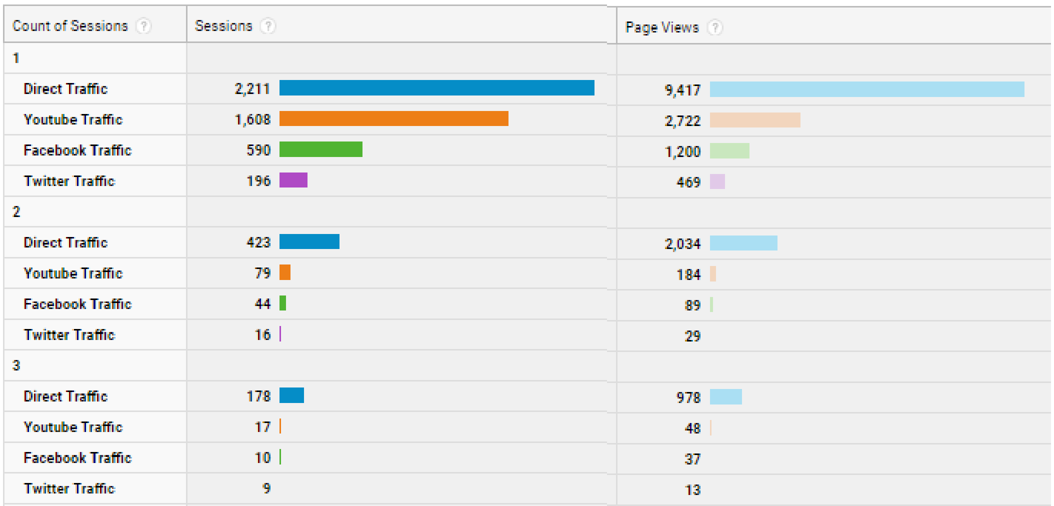
Figure 6.3.: Frequency & Recency Metrics. Source: Screenshot from Google Analytics of `share.catrob.at`

illustrated in Figure 6.3 and Figure 6.4. Therefore, the removal of visitors with one session refines the report and shifts the focus to visitors with two or more sessions [Lew13]. Based on the session count, the audience can be segmented to gather information about the visitor behavior and the engagement value of each segment [Has12; Ken+11; BG15]. Figure 6.4 displays the visitor loyalty of different traffic sources and lists the number of sessions for each segment. The segments are the direct traffic and the social networks, *YouTube*, *Facebook*, and *Twitter*. Hassler [Has12] proposed a segmentation of a website's audience based on the count of session metric. However, the segmentation should be individualized for each website.

- One-time visitors (1 session)
- Interested visitors (2 to 3 sessions)
- Loyal visitors ($>$ 4 sessions)

6.3.2. Days Since the Last Visit

The *Days Since Last Visit* report displays the number of users segmented by the days elapsed since the last interaction with the website took place [Lew13; BG15]. Like the *Count of Sessions* report, the *Days Since the Last Session* report can be used to segment the audience in different categories based on the



Count of Sessions	Sessions	Page Views
1		
Direct Traffic	2,211	9,417
Youtube Traffic	1,608	2,722
Facebook Traffic	590	1,200
Twitter Traffic	196	469
2		
Direct Traffic	423	2,034
Youtube Traffic	79	184
Facebook Traffic	44	89
Twitter Traffic	16	29
3		
Direct Traffic	178	978
Youtube Traffic	17	48
Facebook Traffic	10	37
Twitter Traffic	9	13

Figure 6.4.: Visitor Loyalty Metrics segmented by Direct Traffic and Social Networks. Source: Screenshot from Google Analytics of `share.catrob.at`

visitor recency [Has12; BG15]. Hassler [Has12] proposed a segmentation of a website's audience based on the visitor frequency for the segmentation of the visitors. However, like the segmentation based on the count of sessions metrics, the segmentation based on the day since the last visit metric should be individualized:

- Daily visitors (visitor frequency < 1 day)
- Regular visitor (visitor frequency 1 to 3 days)
- Occasional visitors (visitor frequency > 3 days)

6.4. Geographic Location

Web analytics tools are able to approximately determine the audience's geographic distribution. This technique is called "*Geolocation*" and is done by mapping the address space of the IP-address to a geographic address [HFC11]. The geolocation is able to identify the continent, the country, and the city of the visitor based on the IP-address. Research done by *Cooperative Association for Internet Data Analysis* in the year 2011 [HFC11] analyzed the accuracy of geolocation services, such as *IPligence*[28], *cyScape Inc*[29], and *MaxMind GeoIP*[30], which resulted that 94.5% of the analyzed IP-addresses where successfully mapped to a country. The data accuracy of cities is 76%, therefore, the data accuracy from cities is lower than the data accuracy of countries [HFC11]. Figure 6.5 displays the data accuracy of the *MaxMind GeoIP* service for countries.

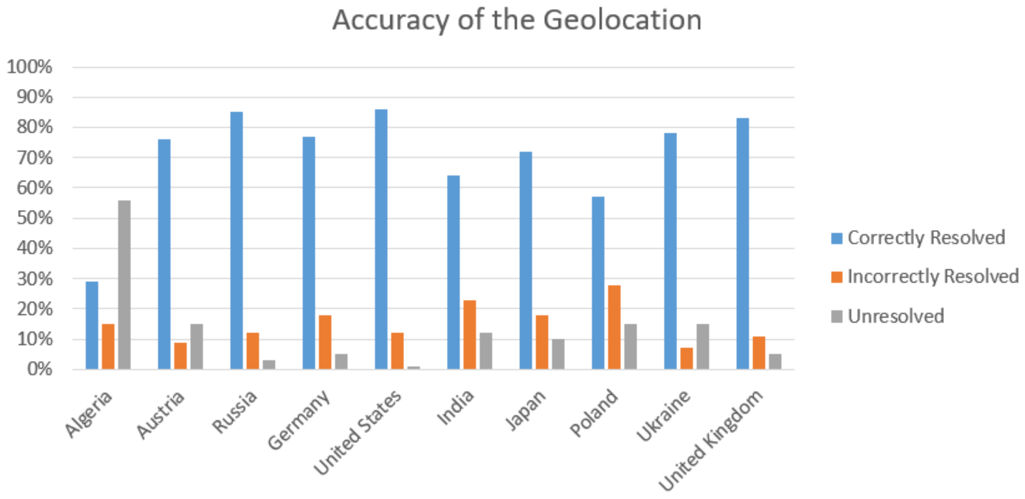


Figure 6.5.: Geolocation Data Accuracy of Countries. Source: MaxMind [Max]

The geographic location is used to give an overview about the geographic distribution of a website's or application's audience. Furthermore, it is possible to individualize marketing or advertising campaigns to target specific countries with specialized advertisements, or to segment the audience based on the geolocation [Web13b; Web13a; McF05; KR14].

[28] www.ipligence.com/

[29] www.cyscape.com/

[30] www.maxmind.com/

6.5. Devices

Website visitors can use different devices to visit a website, such as *tablets*, *smartphones*, or *desktop computers*. Since the introduction of mobile devices and subsequently the development of mobile browsers, web analytics tools are able to gather the same information by page-tagging from mobile devices as from desktop computers [Kau10]. However, the user experience of websites depends on the device [Jan09; Kau10; Bea13]. Therefore, the segmentation based on the device observes these difference and provides information about the user experience for each segment [Bea13]. Furthermore, users are segmentable by details of the device, such as the browser type, the browser version, the screen resolution, or the device type [Web13a].

6.6. Length of Visit

The length of visit metric measures the duration of the interaction with the website in a session [BBW+07; Bea13]. This includes the time spent on a single web page - *"Time on Page"* - and the time spent on the entire website - *"Time on Site"*. These metrics are an indication about the interest of the audience in the website [Has12; Bea13]. The length of the visit can also give a suggestion about the engagement of the content, which indicates the popularity of the viewed content [Kau10; Jan09; Has12].

Figure 6.6 illustrates the measurement of the metrics *Time on Site* and *Time on Page*. The sequence of the visitor steps are:

- At 10:00 the visitor arrives at the "home page" of the website. This triggers the page tag (discussed in section 4.2), which sends information about the *page view* to the web analytics tool.
- The visitor views during the same session "page 2" at 10:01 and "page 3" at 10:05. Afterwards, the visitor leaves the website.
- Each of these page requests triggers a *page view* that contains the load time of the web page.

The *Time on Page* is the subtraction of the current page view's timestamp from the next page view's. However, the time spend on the *exit page* (last

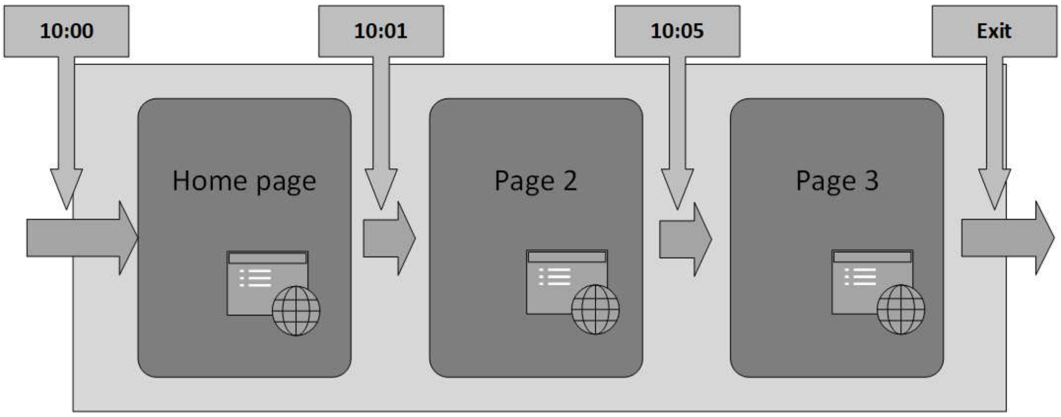
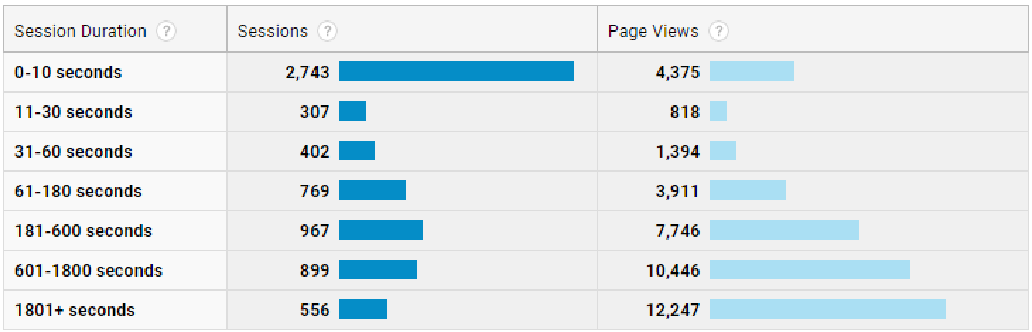


Figure 6.6.: Measurement of *Time on Site* and *Time on Page*. Adapted from [Kau10]

page) cannot be determined because the timestamp after the exit page is not available, as displayed in Figure 6.6 [Bea13; BBW+07]. As a consequence, the time spend on the exit page is set to zero [Bea13; Kau10; BBW+07; Jan09].

The *Time on Site* computation is performed without the consideration of the exit time [Bea13]. This is particularly noteworthy for sessions with only one page view. As a result, the session duration for session with one pageview is zero minutes [Kau10; Bea13; BBW+07].

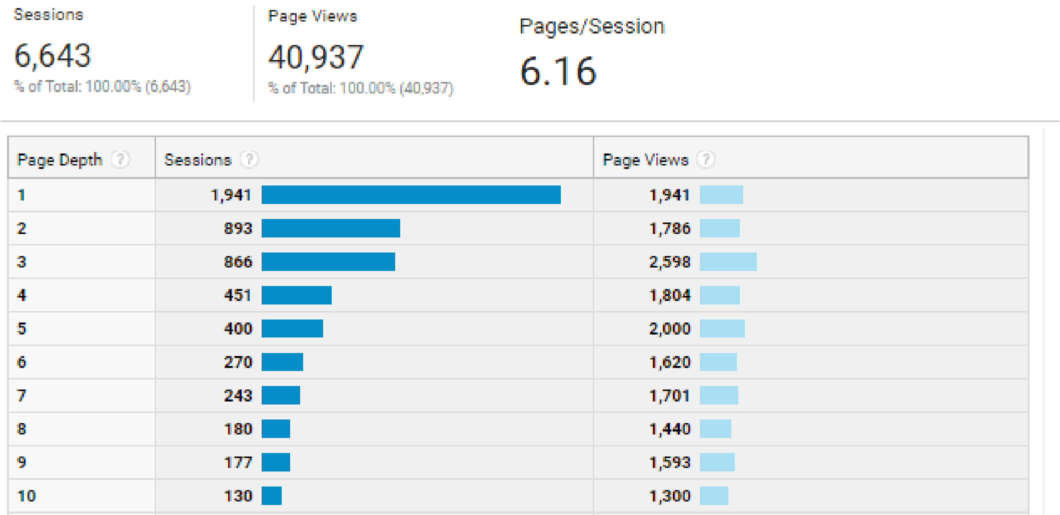


Session Duration	Sessions	Page Views
0-10 seconds	2,743	4,375
11-30 seconds	307	818
31-60 seconds	402	1,394
61-180 seconds	769	3,911
181-600 seconds	967	7,746
601-1800 seconds	899	10,446
1801+ seconds	556	12,247

Figure 6.7.: Distribution of the Session Duration. Source: Screenshot from Google Analytics of `share.catrob.at`

In web analytics tools the *Time on Site* and the *Time on Page* metrics are displayed as an average value, as illustrated in Figure 6.1. However, the average value does not provide significant information about the user behavior [Kau10]. The distribution of the length of the visit, shown in Figure 6.7, provides essential information about changes in the user behavior [Kau10; Has12]. For example, the performance of website modifications is observable by the visitor distribution, or user satisfaction by monitoring the overall session duration distribution [Has12; Kau10]. Additionally, the session duration is used to segment the audience to investigate the performance of traffic sources and search keywords [Has12; Bea13; Jan09]

6.7. Depth of Visit / Pages per Visitor



Sessions
6,643
% of Total: 100.00% (6,643)

Page Views
40,937
% of Total: 100.00% (40,937)

Pages/Session
6.16

Page Depth	Sessions	Page Views
1	1,941	1,941
2	893	1,786
3	866	2,598
4	451	1,804
5	400	2,000
6	270	1,620
7	243	1,701
8	180	1,440
9	177	1,593
10	130	1,300

Figure 6.8.: Page Depth Distribution. Source: Screenshot from Google Analytics of `share.catrob.at`

The *Depth of Visit* metric measures the average number of page views per session [BBW+07], as shown in Figure 6.8. The metrics is similar to the *Length of Visit* metrics and indicates the visitor's interest in the offered content of the website [Has12; PPC12; Kau10; BG15]. Figure 6.8 displays the distribution of sessions based on the *Page Depth* and displays the number of pages viewed during a session. The comparison of the distribution of different segments provides information about their engagement value [Has12; Bea13].

6.8. Bounce Rate

The *Bounce Rate* metric is a special kind of the *Depth of Visit* metric and displays only visits with a page depth of *one* [Has12; PPC12; Bea13]. Web pages with a high bounce rate do not animate the visitor's interest in the website [Has12; Kau10] and indicate either a not engaging website design or content [HMP09]. Web analytics tools are able to measure the bounce rate of a website on two different levels: the *entire website* or *individual web pages*. Based on these two levels, the engagement value of the entire website or individual web pages is ascertainable [Has12]. Furthermore, by observing the bounce rate of *referring websites*, *search keywords*, or *paid keywords* is their performance measurable [Kau10; Bea13; Cli12]. Especially after SEO optimizations or advertisement campaigns, the monitoring of the bounce rate reveals the success of the optimization or campaign [Has12; Ken+11].

The bounce rate metric is not able to provide meaningful information for single page websites or web pages with high interactive content, because the standard page tag gathers data about page views, not interactions with the web page [Has12; Cli12]. Additional tracking code, such as the event tracking or virtual pageviews, as discussed in chapter 7, is required to track on-site interactions.

6.9. Visitor Clickstream and Behavior Flow

The behavior flow report observes the visitor's path through the entire visit and examines the visitor journey from the landing page to the exit page [Kau07; Has12; Jan09]. Figure 6.9 displays such a behavior flow graph beginning from the landing page, the navigation, and the exit page. In an article of WAA, two concepts of visitor behavior were introduced [Jan09; BJ10]:

- the goal-driven behavior
- the random and illogical behavior

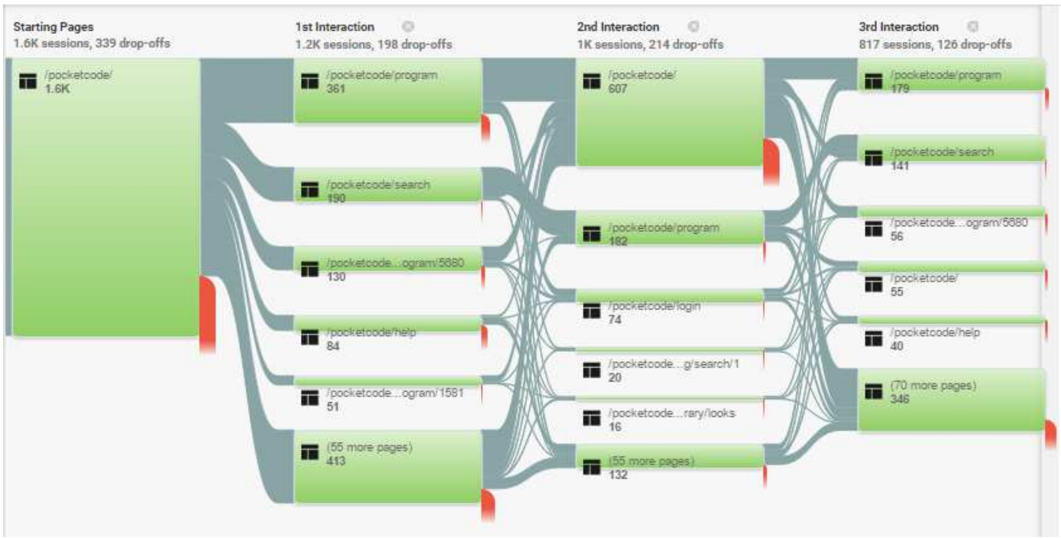


Figure 6.9.: Behavior Flow Graph. Source: Screenshot from Google Analytics of `share.catrob.at`

In the goal-driven behavior, the visitor follows a logical and linear path [Jan09] from the landing page to the goal. Any exit from this path implies a visitor's confusion and subsequently an occurring problem [Jan09; Kau07]. The second random and illogical behavior assumes that the user behavior of the visitor is only influenced by the currently viewed web page [Jan09; BJ10].

However, the complexity of the path analysis increases exponentially with the number of web pages and the observed steps [Kau07; Cli12; Has07]. Furthermore, only a small amount of visitors are using the same sequence of steps through the entire website [Kau07; Cli12]. As a result, an analysis of a specific click-path represents only a minority of visitors [Kau07; Cli12; Bea13]. The path analysis is only meaningful for websites with a small number of web pages or for websites with grouped web pages (content grouping) [Kau07; Cli12; Bea13; Has12]. The content grouping reduces the number of web pages through a grouping of similar web pages [Has12; Goo19c].

An in-depth path analysis is not meaningful because of the underrepresentation of visitors, because of the various paths a visitor can take from one web page [Kau07; Bea13; Cli12]. The click-path analysis observes the relations between web pages and investigates the *incoming and outgoing traffic* of individual web pages or groups of web pages [Bea13; Cli12].

A further method to analyze the visitor path is the *funnel analysis*. The funnel analysis examines, in opposition to the behavior flow, only a *single path* of the visitor to a predetermined goal [Has12; Cli12]. Any *derivation* from this path indicates confusions or problems [Has12; Cli12]. The gained knowledge of the funnel analysis helps to optimize this visitor path to reach the goal [Has12].

6.10. Internal Search

Most of the metrics are providing data based on numbers or ratios without any interpretation of the results or insights. In opposition, an analysis of the internal search keywords provides insights about a user's intent [Kau10; Has07] and search trends [Ald06]. Furthermore, the internal search analysis supports the determination of successful searches and unsuccessful searches (bounce rate), or hard to find web pages with their respective keywords [Ogl10; Has07]. This information can be used to provide missing content for the visitors or to add new features to the website [Ogl10; Ald06]. The internal search also reveals the most frequently used search keywords, the usage frequency, and the user behavior after a search (time on site, further searches, and navigation path) [Has12; Kau10; Cli12].

6.11. Page Views

The page view metric provides information regarding the frequency of viewed web pages and the general interest of visitors [Kau10; Has12]. Web pages with the most page views give an indication about the content visitors are interested [BJ10; Bea13]. Furthermore, a website's navigation should be optimized to reach web pages that are aligned with the business goals of the website [BJ10; Has07]. Volatile changes in the page view count signify changes in visitor trends [Kau07; Kau10; Has12]. For example, advertisements, new life trends, or new referrals are causes for these changes [Kau07;

Haso7]. The page view metric also identifies web pages with low popularity [Kau07; Has07].

6.12. Landing Page

The landing page is the first web page a visitor arrives at the website [BBW+07]. Since a high amount of traffic origins from organic search traffic [Zec14] and search engines not always link to the homepage, the landing page can differ from the intended entry page of the website [BBW+07; Bea13]. The top landing pages metric serves the purpose to identify these mostly used landing pages. The bounce rate is an indication of the visitor's engagement to measure the performance of these landing pages and there respectively traffic sources and search keywords [Kau10; Jan09; Cli12]. The landing page affects the user behavior of the visitor [Bea13], therefore an optimization influences the overall performance of the website [BJ10].

6.13. Exit Page

The exit page is the last page the visitor views during a visit [BBW+07; Cli12]. An analysis of the exit page identifies the success of a user journey [Bea13; Has07]. An in-depth investigation of the visit is necessary to determine if the last interaction was an exit or an abort of the session [Kau07; Bea13; Has07]. Kaushik [Kau07] recommends only an investigation of the exit by *processes of closed nature*, for example, the checkout process, or a multi-page registration process to identify points of confusion or missing content [BJ10; Jan09].

6.14. Site Speed

The site speed report measures the time needed for a full web page load. The monitoring of the site speed is of great importance for the visitor engagement and loyalty [Jan09; Cli12]. The study *"Consumer Response to Travel Site Performance"* by *Akamai Technologies, Inc* did reveal that 57% of the visitors will leave the website if the page load time is greater than two to three seconds [Aka10]. Younger people are even less patient and will abandon the website even earlier. An improvement of the page load time can consequently increase the revisiting value by raising visitor loyalty and decreasing the bounce rate, caused by a slow web page load.

6.15. Video Tracking

Video tracking is part of the event tracking and can be used to evaluate the performance of the embedded video content of a website. Capturing information of the video content is similar to the tracking of user-interaction on RIAs and is implemented through customized event tracking code. The obtained tracking data is used to measure the performance of the video content on two levels: [Has12]

- General Performance of the Video Content
- Individual Performance of Videos

An indication of the performance of a website's video content yields the evaluation and interpretation of the following metrics identified by Kaushik [Kau10] and Hassler [Has12]:

- ***Play rate***
  The play rate is the number of videos played per unique visitor. This metrics represents how engaging the video content of the website is for website visitors [Eng16]. A high play rate indicates either that the video content is highly accepted by the audience [Has12; Kau10; Eng16] or that the content of the video is difficult to comprehend [Has12].
- ***Abort rate***
  The abort rate is the number of video starts subtracted by the number of completed videos per visit. The metric provides information about the interests of the audience [Has12; Kau10]. Subsequently, a high abort rate indicates that the content of the videos does not match with the visitor interests [Has12; Kau10].

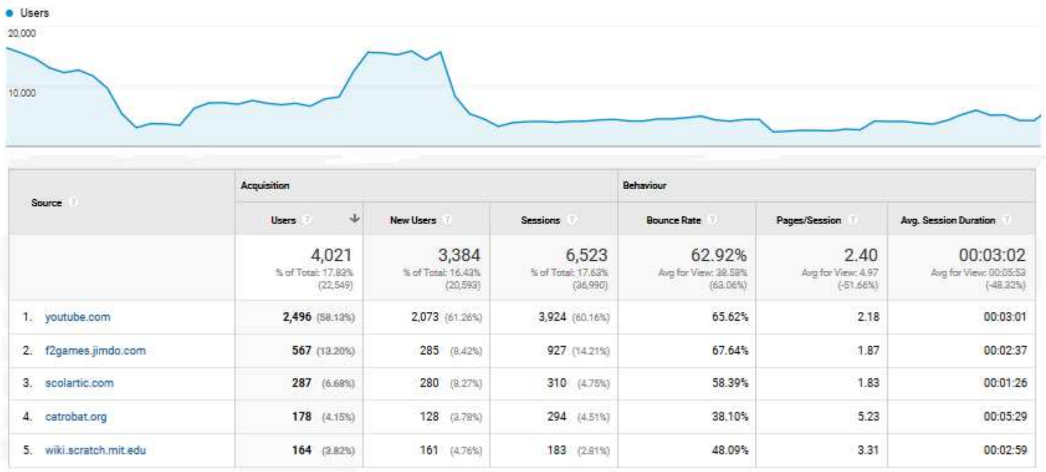
- ***Average play time***
  The average play time measures the time spent watching the video content. Underperforming videos are identifiable by comparing the individual play time to the global avg. play time [Has12].
- ***Engagement behavior***
  The engagement behavior is compromised of the number of re-watched scenes of a video, the skipped scenes of the video and the viewed length of a video, which can be measured in units of 25%, 50% 75% and 100% [Eng16; Has12; Kau10]. This information is used to evaluate the quality of the videos [Eng16]. Significant peaks in re-watched scenes or in the scene where the video is aborted either indicates confusing video content or highly engaging content [Has12; Kau10]

6.16. Direct Traffic

Traffic identified in web analytic tools as direct traffic can originate from the direct input of the URL, browser bookmarks, saved web links [Kau10; PPC12; Has07], or browser autocompleted URLs [Bea13]. However, the different direct traffic sources are not distinguishable [Has12]. The direct traffic is an indication for the visitor loyalty [Kau10; Has12; Pra+13]. Based on the traffic source, differences in the user behavior can be identified [Pra+13; Ken+11]. The direct traffic metric responds slowly to irregular visitor changes [Kau10]. Subsequently, only long-term developments in visitor trends are observable [Kau10]. A decreasing trend in the direct traffic metric indicates a loss of interest in the content of the website [Has12]. Furthermore, the direct traffic metric provide information about the success of offline advertisement campaigns by monitoring the sessions changes during a campaign [Has12; Jan09].

6.17. Referral Traffic

Referral traffic originates from third-party websites by interacting with a hyperlink or an event that redirects the visitor to the observed website [Jan09;



Source	Acquisition			Behaviour		
	Users	New Users	Sessions	Bounce Rate	Pages/Session	Avg. Session Duration
	4,021 % of Total: 17.83% (22,549)	3,384 % of Total: 16.43% (20,593)	6,523 % of Total: 17.63% (36,990)	62.92% Avg for View: 38.58% (63.06%)	2.40 Avg for View: 4.97 (-51.66%)	00:03:02 Avg for View: 00:05:53 (-48.32%)
1. youtube.com	2,496 (58.13%)	2,073 (61.26%)	3,924 (60.16%)	65.62%	2.18	00:03:01
2. f2games.jimdo.com	567 (13.20%)	285 (8.42%)	927 (14.21%)	67.64%	1.87	00:02:37
3. scolartic.com	287 (6.68%)	280 (8.27%)	310 (4.75%)	58.39%	1.83	00:01:26
4. catrobat.org	178 (4.15%)	128 (3.78%)	294 (4.51%)	38.10%	5.23	00:05:29
5. wiki.scratch.mit.edu	164 (3.82%)	161 (4.76%)	183 (2.81%)	48.09%	3.31	00:02:59

Figure 6.10.: Referring Traffic Source Domains. Source Screenshot from Google Analytics of `share.catrob.at`

BBW+07]. The referral traffic metric provides information about the last visited website of a user. A content analysis of the referring web pages provides insights about visitor expectations [Kau10; Has12; Bea13]. Subsequently, the bounce rate metric and time on site metric contributes further information to deduce the fulfillment of the visitor expectations [Kau10; Has12]. A low bounce rate indicates a high affinity for the provided content. The information of the highest referring website domains, illustrated in Figure 6.10, identifies websites that are generating a significant amount of traffic. In opposition to the direct traffic metric, the referral traffic metric responds accurately to short time trends [Has12].

6.17.1. The Long Tail - Pareto Principle

The *Pareto Principle* by *Vilfredo Pareto* describes a reoccurring phenomenon based on the 80/20 rule [San87]. This principle explains the natural occurrences of 80/20 ratios, for example, in e-commerce: 20% of the products are generating 80% of the total sales [SA08]. The *Pareto Principle* can be applied to numerous areas, such as sociology, economy, politics, information

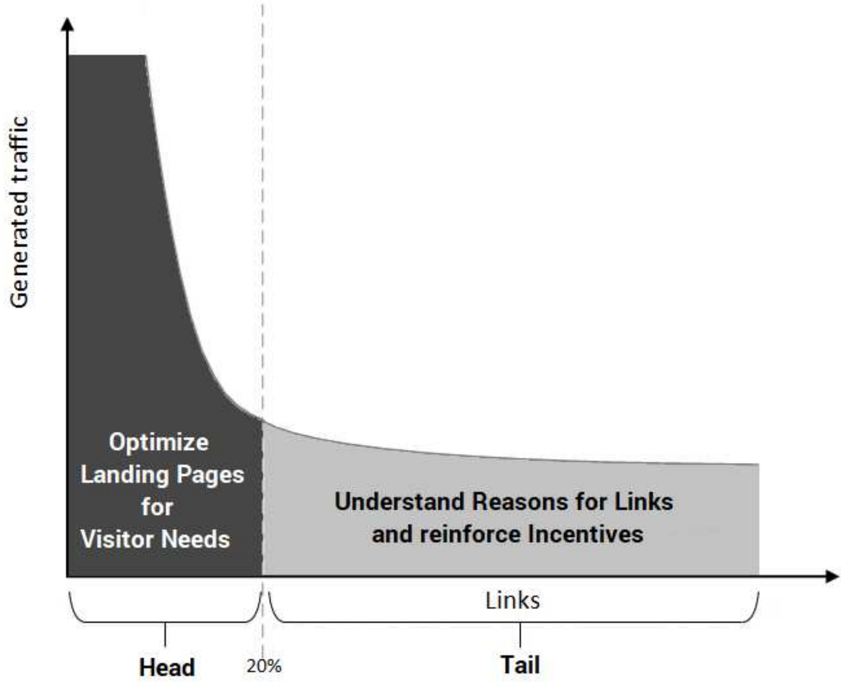


Figure 6.11.: Long Tail of Referring Links. Adapted from [Has12]

technology, and so forth [San87]. The *Long Tail* introduced in the book *"The Long Tail: How Endless Choice is Creating Unlimited Demand"* by Anderson [And06] applies this *Pareto Principle* to e-commerce. In his book, Anderson advises to also concentrate on the 80% of products (The Tail) that are only generating 20% of sales. Even if the products do not generate a high amount of revenue, they attract potential customers to the platform [SA08]. This serves the purpose to increase the sales of the 20% products (The Head) through the 80% products (The Tail) [BHS11].

The *Pareto Principle* can also be used for the analysis of the referral traffic sources by interpreting the referral traffic metrics. Figure 6.11 shows the adapted version of the *Long Tail* by Hassler [Has12] and provides recommendations for the traffic generating links. For high traffic generating links - *The Head* - it is recommended to analyze the visitor expectations of each referring website and to optimize the landing page for the visitor needs. The metrics of the behavior analysis (discussed in section 5.3.1) especially the bounce rate metric provides an indication for the performance of the landing pages in regards to referring websites [Kau10; Has12]. For links that are generating a low amount of traffic - *The Tail* - it is recommended to analyze the reason why third-party websites referred to the website [End+08; Has12].

6.18. Search Engine Traffic

Web analytics tools are categorizing traffic from search engines, such as Google, Bing, Yahoo! Search, DuckDuckGo, or Ask as organic search traffic. According to a study from *"Search Engine Watch"*[31] search engine traffic is responsible for approximately 64% of the total website traffic, which is illustrated in Figure 6.12 [Zec14]. Furthermore, the study did state that 60% of the identified direct traffic is misidentified. This misidentified traffic is mainly traffic from search engines. This reaffirms that search engine traffic is the most important traffic source for websites.

[31] `https://searchenginewatch.com`

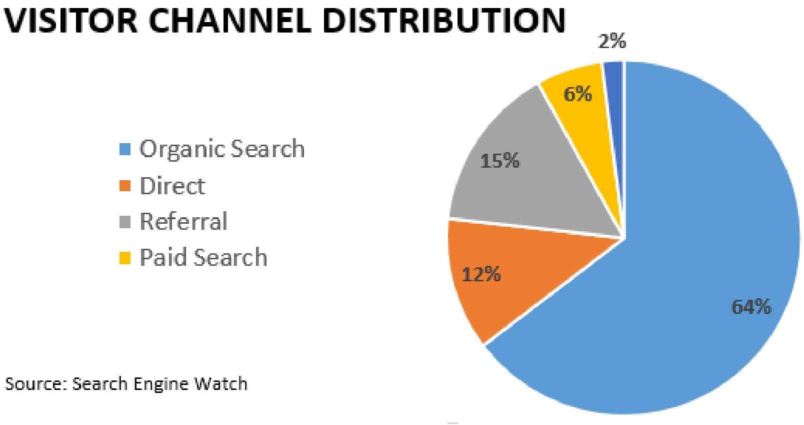


Figure 6.12.: Visitor Channel Distribution. Adapted from [Zec14]

Web analytics tools are able to identify and segment the search engine traffic by the used search engine. This information supports the identification and rating of individual search engines. Moreover, the segmented traffic can be used to optimize the website for underperforming search engines (Search Engine Optimization) [Has12]. The *Search Engine Optimization (SEO)* is used to reach a higher *page rank*, and consequently, an increase in the generated website traffic [Zil15; DR11]. A comparison of the website traffic before and after a website optimization measures the performance of the performed SEO [Zil15]. Additionally, the metric provides information about the indexation status of a website, for example, if no web pages or wrong web pages are indexed [Has12; DR11]. An incorrect indexed website can have multiple reasons, such as the website does not allow the crawling of the website (for example, using *robots files*[32] or *meta tags*[33] to disallow the crawling of search engines), or the crawler is not able to process the website content because the website is solely implemented in JavaScript, as discussed in chapter 4.

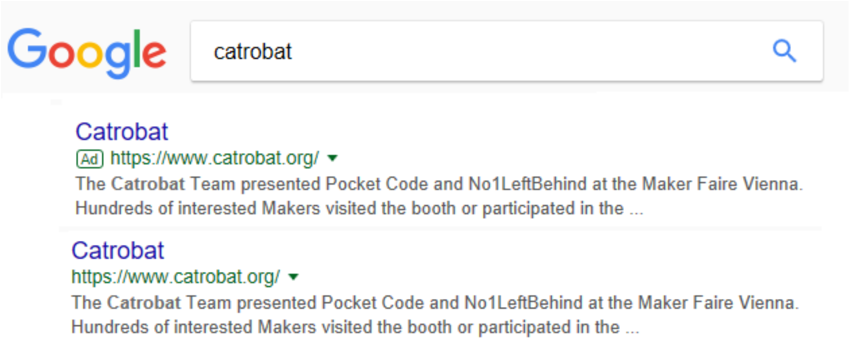
[32] `http://www.robotstxt.org/`

[33] `https://developer.mozilla.org/en-US/docs/Web/HTML/Element/meta`

6.19. Search Engine Keywords

Web analytic tools are not only determining the used search engine but also the used keywords to reach the website [BBW+07; Has07]. An analysis of the search keywords provides information about the intent and expectations of the visitors [BBW+07; Kau10]. Furthermore, the expected content for each individual keyword can be inferred [PPC12; Bea13]. The bounce rate metric, depth of visit metric, time on site metric, visitor metric, or conversion metric are measuring the performance of the search keywords [Kau10; Bea13]. Search engine keywords can be differentiated in two types, illustrated in Figure 6.13:

- Organic Search Keywords
- Paid Search Keywords



Catrobat
Ad https://www.catrobat.org/ ▾
The Catrobat Team presented Pocket Code and No1LeftBehind at the Maker Faire Vienna. Hundreds of interested Makers visited the booth or participated in the ...

Catrobat
https://www.catrobat.org/ ▾
The Catrobat Team presented Pocket Code and No1LeftBehind at the Maker Faire Vienna. Hundreds of interested Makers visited the booth or participated in the ...

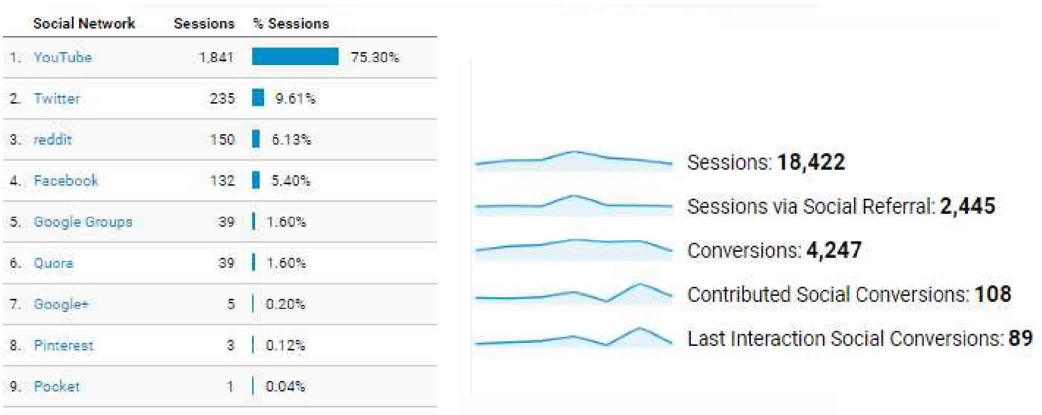
Figure 6.13.: Organic and Paid Search Results. Source: Screenshot Google (modified)

Organic Keywords Search results from organic search are generally without a particular prefix in the search results, as illustrated in the second entry in Figure 6.13. The organic search keyword report provides for each individual keyword information about the *acquisition*, such as the number of sessions, the percent of new sessions, or the number of new visitors. The report also includes data about the *visitor's behavior*, such as the bounce rate, the visited pages per session, or the average session duration, discussed in section 5.3.1.

Paid Keywords Search results from paid keywords are usually marked with an advertisement identifier. This can be an *"Ad"* prefix in front of the URL, as illustrated in the first entry in Figure 6.13. Paid search results are either placed on top of the organic search results, or beside the search results [Has12; DR11].

A division in paid and organic search keywords prevents the mutual influence of the metrics to measure the individual performance the search results [Kau10; Bea13]. Web analytics tools are able to integrate data into the analytic tool directly, for example, AdWords[34]. A further application for *The Long Tail*, discussed in section 6.17, is the analysis of the performance of search engine keywords [Has12; DR11; Bea13; Zil15]. Furthermore, the landing page must reflect the visitor expectations to increase the performance of the search keywords [Kau10; Has12]. By monitoring the changes in the keyword ranking, seasonal trends and the influence of marketing campaigns are observable [Has12].

6.20. Social Media Traffic



	Social Network	Sessions	% Sessions
1.	YouTube	1,841	75.30%
2.	Twitter	235	9.61%
3.	reddit	150	6.13%
4.	Facebook	132	5.40%
5.	Google Groups	39	1.60%
6.	Quora	39	1.60%
7.	Google+	5	0.20%
8.	Pinterest	3	0.12%
9.	Pocket	1	0.04%

Sessions: **18,422**

Sessions via Social Referral: **2,445**

Conversions: **4,247**

Contributed Social Conversions: **108**

Last Interaction Social Conversions: **89**

Figure 6.14.: Social Media Traffic Overview. Source: Screenshot from Google Analytics of `share.catrob.at`

[34] `https://adwords.google.com/`

The social media traffic report, illustrated in Figure 6.14, gives an overview of the social media services that are generating traffic for the website. Social media networks are channels to directly engage with the audience [HRC11]. The direct interaction with visitors strengthens the customer relationship and provides immediate feedback to improve the website [Naj05].

7. Event Tracking

In the previous chapter, the web reports and metrics were discussed, that are gathering information by tagging the web page with the standard page tag. Since the page tag only sends data during a full page load [Goo18], user interactions are not tracked that do not execute a full page load. This can either be *outgoing interaction* (for example, downloads or outgoing links), or *on-site interactions* (for example, the number of social interactions, AJAX interactions, or videos statistics) [Goo19d]. *Virtual pageviews tracking* or *event tracking* is able to capture such on-site interactions [Cli12; Bea13].

Rich Internet Applications Like previously stated, the standard tracking code is not able to track interactions that do not execute a full page load. For *Rich Internet Applications (RIA's)*, that mostly consist of highly interactive and dynamically loaded content [NH05], reduces this the traceable content and valuable information about the website interaction is lost [Cli12].

7.1. Virtual Pageviews

Virtual pageviews are similar to standard page views. However, instead of triggering during a full page load, virtual pageviews are triggered by user interactions or events. These user interactions or events can be a click on a hyperlink, picture, button, the scrolling of the web page, or the loading of dynamic content [Has12]. The tracking code of the page tag must be modified to track these actions [Cli12; Goo18]. Sharma [Sha19b], Clifton [Cli12], and Kaushik [Kau10] recommend to track any interaction that corresponds to a page view by virtual pageviews and every other interaction through event tracking.

Since each virtual pageview equals a standard page view, the virtual pageview increases the total number of page views artificially [Cli12]. Virtual pageviews are an essential part of the funnel analysis, discussed in section 6.9, to track the completed steps of a web form or the executed steps in a shopping process [Sha19b].

```
// Virtual Pageview Syntax
gtag(<command> (’config’), ’GA_TRACKING_ID’, {
    ’page_title’: (document.title),
        // ’Virtual Pageview’
    ’page_path’: ’/virtual’ + (document.location.pathname),
        // ’/virtual/example/new-page.html’,
    ’page_location’: (document.location)
        // ’https://www.example.com/example’
    });
```

Listing 7.1: Virtual Pageview Tracking Code for Google Analytics Page Tagging

Listing 7.1 shows an example code for the virtual pageview tracking code in Google Analytics. The virtual pageview tracking code includes the parameters[35]:

- **command**
  The command parameter specifies the type of the command and can be set to *config*, *set*, or *event*. The config value determines the send *event* as page view.
- **GA_TRACKING_ID**
  The GA_TRACKING_ID must be replaced by an individual tracking ID to identify the send events to their corresponding website.
- **page_title**
  The HTML-page title[36] of the currently viewed HTML-document[37] is set by the title-tag in the header of the HTML-document. The default value of the title parameter is the DOM-element `document.title`, for

[35] https://developers.google.com/gtagjs/reference/api

[36] https://developer.mozilla.org/en-US/docs/Web/HTML/Element/title

[37] https://www.w3schools.com/jsref/dom_obj_document.asp

example, for the homepage of the Catrobat community website[38] the document title is "Pocket Code Website".

- **page_path**
  The page_path parameter specifies the pathname of the generated pageview. Since web analytic tools do not differentiate between standard page views and virtual pageviews, it is necessary to include an identifier in the path parameter to identify them in the reports [Sha19b]. For example, `/virtual/virtualpageview.pdf` for the download of a PDF-document. The keyword `/virtual/...` differentiate the virtual pageview from a standard pageview.
- **page_location**
  The value of the page_location parameter can be obtained by the DOM-element `document.location`[39] and includes the entire URL of the currently viewed HTML-document. For example, `https://share.catrob.at/pocketcode/` for the homepage of the Catrobat community website.

7.2. Event Tracking

Event tracking similar to the tracking of virtual pageviews is able to track user-interactions that do not require a full page load [Cli12; Has12]. In opposition to virtual pageviews, event tracking is not considered as a page view by web analytic tools and does not increase the page view metric artificially [Cli12]. Event tracking should be used for user-interactions that do not correspond to page views [Sha19a; Cli12; Kau10].

Event tracking can be used to track [Sha19a; Has12]:

- Videos
- Gadgets
- Podcasts
- Images
- Buttons

[38] `https://share.catrob.at/pocketcode/`

[39] `https://www.w3schools.com/jsref/obj_location.asp`

- Forms
- Scroll bars
- External links
- Light box
- AJAX content
- Mouse movement

```
// Event Tracking Syntax:
gtag(<command>, <event_action>, {
  ’event_category’: <category>,
  ’event_label’: <label>,
  ’value’: <value>
});

gtag(’event’, ’play’, {
  ’event_category’: ’Videos’,
  ’event_label’: ’Tutorial 1’,
  ’value’: 1,
  ’noneInteraction’: false
});
```

Listing 7.2: Event Tracking Code for Google Analytics Event Tracking

Listing 7.2 shows an example code for the event tracking code in Google Analytics. The event tracking code includes the following parameters[40]:

- **command**
  As in section 7.1 introduced, the command parameter specifies the type of the command and can be set to *config*, *set*, or *event*. The command parameter must contain the value *event* to determine the send event as part of the event tracking.
- **event_category**
  The event_category can be used to group events of similar types. For example, YouTube videos, outgoing links, internal clicks, downloads et. cetera.

[40] https://developers.google.com/analytics/devguides/collection/analyticsjs/field-reference

- **event_action**
  The event_action parameter designated the name of the event action. This can be the type of the event or interaction, such as the video *play* or *pause* event [Sha19a], but it can also include additional information, for example, the used video platform or the current URL [Goo19a].

  Duplicate event actions Event action names are globally valid [Goo19a]. This means events with the same event action name are aggregated, even from different event categories [Goo19a; Cli12]. Event action names should easily be related to the performed event but distinct in different event categories [Sha19a]. Due to the global scope of event action names, only the first interaction with an event of the same action name is a *unique action* [Goo19a].
- **event_label**
  The event_label parameter is optional and can be used to describe the content of the tracked event further. For example, for videos the title of the video can be chosen [Has12], or for downloads the filename of the downloaded file [Goo19a].
- **value**
  The optional value parameter can be used to associate a numerical value with the event, which enables the weighting of events [Goo19d]. For example, the rating of social interactions, or the measurement of social engagement, which is discussed in section 7.3 [Has12].
- **noneInteraction**
  The noneInteraction parameter defines the influence of the event on the bounce rate (discussed in section 6.8). Events that are not considered as web page interactions, the noneInteraction parameter should be set to true, for example, the scrolling of the web page.

7.3. Social Engagement

The engagement value of user-generated content, such as videos, blogs, forum posts, or social buttons [KH10], illustrated in Figure 7.1, can be measured by the *Social Engagement Score* [Has12; GH11]. Table 7.1 contains a suggestion to weight the individual social interactions of a website. For

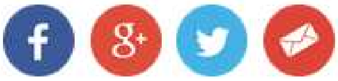


Figure 7.1.: Social Interaction Buttons

example, Facebook Share, Google Share, Twitter Tweets, or Comments to calculate the social engagement score for the content. The calculated social engagement score is then displayed in the social engagement score metric by web analytics tools to quantify the engagement value of the website [GH11]. Listing 7.3 shows an example tracking code to measure the *Social Engagement Score*.

Social interaction	Social Value
Comment	10
Facebook Like	5
Facebook Share	3
Google +1	5
Google Share	3
Content Rating	1-5
Social Bookmark	2

Table 7.1.: Social Values. Adapted from [Has12]

Social interactions are capturing the user interactions of website visitors by tracking the interactions with *social buttons* provided by social networks [GH13]. These can be buttons from social networks like Facebook, Google+, or Twitter. The social interactions provide information to segment the audience based and the social network and the amount of interaction (for example, *Social Actions*, *Unique Social Actions* and *Actions per Social Session*) with those social networks [Cli12; Has07; Goo19b]. Social interactions are encouraging interactions with the content of the website [GH13]. Furthermore, the segmented audience can be used to target specific social networks in marketing campaigns [Cli12]. Listing 7.4 shows the tracking code for social interactions in Google Analytics.

```
// Event Tracking - Social Engagement
gtag(<command>, <event_action>, {
  ’event_category’: <category>,
  ’event_label’: <label>,
  ’value’: <value>
});

gtag(’event’, ’SocialEngagement’, {
  ’event_category’: ’engagement’,
  ’event_label’: document.title,
  ’value’: 3
});
```

Listing 7.3: Social Engagement Score Tracking Code for Google Analytics Event Tracking

```
// Event Tracking - Social Interactions
gtag(’event’, ’share’, {
  ’method’: [socialNetwork]
  ’content_type’: [article, video],
  ’content_id’: [article-ID, video-ID]
});

gtag(’event’, ’share’, {
  ’method’: ’Facebook’
  ’content_type’: ’video’,
  ’content_id’: ’video-5555’
});
```

Listing 7.4: Social Interaction Tracking Code Google Analytics

8. Catrobat

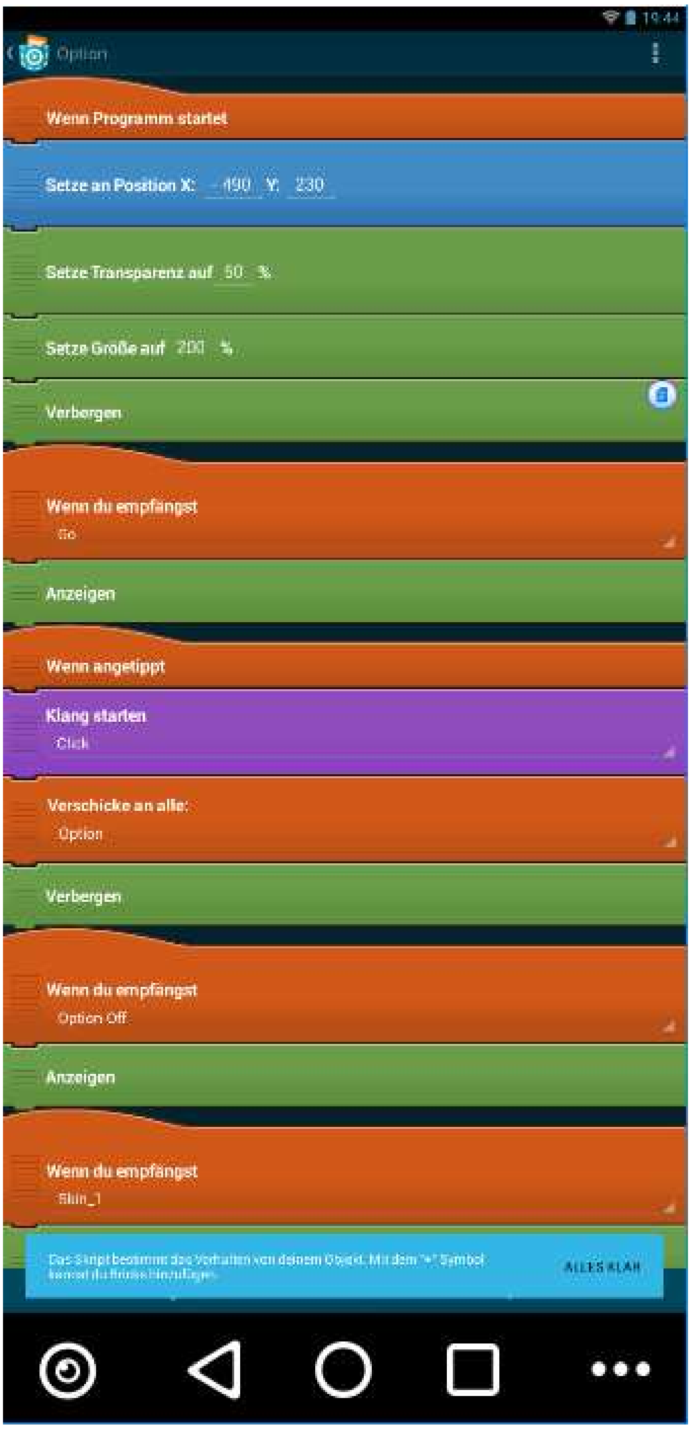
In the previous chapters, the theoretical background of web analytics was discussed. Introduced was the technology to gather information about website visitors, the log file analysis and the page tagging (discussed in chapter 4). Furthermore, differences between off-site and on-site metrics were discussed in chapter 5.

8.1. Catrobat

In the subsequent chapters, the knowledge from the previous chapters will be applied for the analysis of the Catrobat community website's visitors. The Catrobat sharing website that is available at `https://share.catrob.at/pocketcode/`, is a subproject of the Catrobat project `https://www.catrobat.org/` and Google Analytics[41] will be used to analyze the user behavior of the visitors. The Catrobat project is a non-profit open source project that has the vision to allow every person of humanity, free of origin, wealth, or gender to acquire computational thinking skills [Sla14]. Catrobat provides for this vision an environment for children and teenagers to create free educational applications and offers different development environments for the creations of the Pocket Code programs [Cat19]. The targeted platform of the development environments are smartphones. Supported operation systems (OS) for these development environments are Android, and the development environments for iOS and HTML5 capable browsers are in the beta stage [Sla14].

Catrobat is inspired by the *Scratch programming language*, available at `https://scratch.mit.edu/`, which was developed by the *Lifelong Kindergarten*

[41] `analytics.google.com`



Wenn Programm startet

Setze an Position X: -190 Y: 230

Setze Transparenz auf 50 %

Setze Größe auf 200 %

Verbergen

Wenn du empfängst
Go

Anzeigen

Wenn angetippt

Klang starten
Click

Verschicke an alle:
Option

Verbergen

Wenn du empfängst
Option Off

Anzeigen

Wenn du empfängst
Skin_1

Das Skript bestimmt das Verhalten von deinem Objekt. Mit dem "+" Symbol kannst du Bricks hinzufügen.

Figure 8.1.: Pocket Code - Lego Block Style

Group at the MIT Media Lab. In opposition to Scratch, Catrobat targets mobile devices and does not require a desktop computer for the development and execution of the programs [Har+13]. Pocket Code is an on-device visual programming system and supports the creation and execution of applications. The programming language of Pocket Code is similar to Scratch and resembles a LEGO block-like building concept [Sla14], as shown in Figure 8.1.

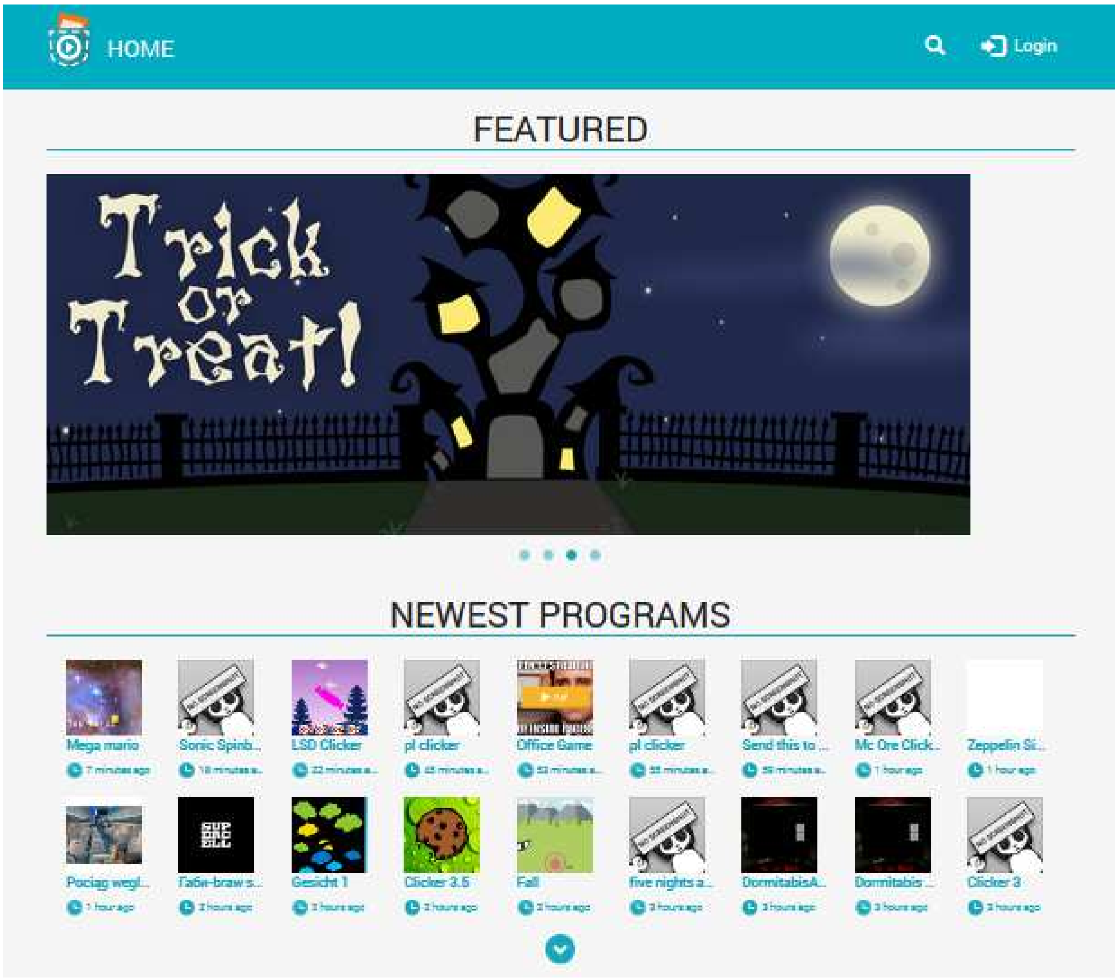


Figure 8.2.: Catrobat - Community Website

The sharing website, illustrated in Figure 8.2, is used to share the created

Pocket Code programs with the entire community. The purpose of the sharing website is to provide every user of the Pocket Code community the ability to download existing programs and subsequently to modify them. This is called remixing and is an essential feature of the Pocket Code community website. Remixed programs can be re-uploaded to the sharing website to be re-shared with the Pocket Code community. [Sla12]

Additionally, Catrobat assists in the development of creative content with the Pocket Paint[42] and the Musicdroid application. Pocket Paint is an image manipulation program that supports the creation and modification of pictures. Musicdroid is a currently developed but not released extension of Pocket Code and provides an environment to create music.

8.2. Target Group

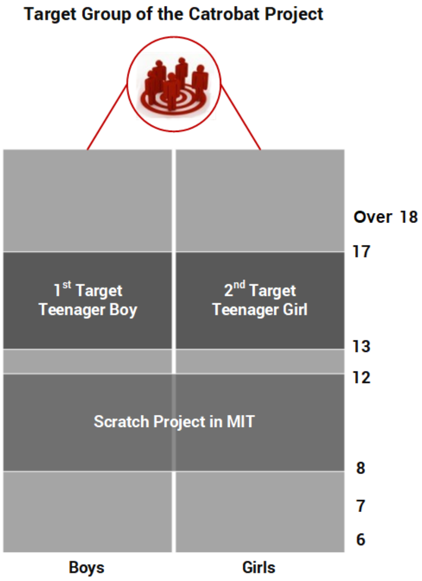


Figure 8.3.: Pocket Code - Target Group. Adapted from [Tan12]

[42] https://github.com/Catrobat/Paintroid

The target group of Pocket Code is children and teenagers of the age between 13 to 17 years, as shown in Figure 8.3 [Tan12]. In comparison to Scratch, that targets children between the age of 8 and 12, Pocket Code focuses on slightly more grown-up teenagers that are using mobile devices. However, it is possible that even children under the age of 13 years use the community website or application. Since the targeted audience are minors, it is especially important to protect the privacy of the website visitors, which is further discussed in chapter 10.

8.3. Catrobat Community Website

The Catrobat community website, illustrated in Figure 8.2, was implemented to provide the users and developers of the Pocket Code application a platform to share their programs. Every member of the community is able to *Like* and *Comment* shared programs. By providing tutorials and video guides, the sharing website offers support for users in the program creation. The website is available in 53 different languages to enable a wide audience access to the community website. The Pocket Code application[43] is the main development environment for Pocket Code programs. This Pocket Code application provides features for the developing, downloading, and sharing of the created programs. Additionally, the Pocket Code application provides resources to assist the developer in the program creation. The application used the an integrated browser to access resources of the Catrobat community website. Since the application uses the integrated browser for the website access, interactions from the Pocket Code application are identifiable by the User-Agent of the HTTP-header and can be tracked by web analytics tools.

The community website consists of six main web page categories: the homepage, the program page, the profile page, the search page, the help section, and the media library.

The *homepage* contains a collection of featured programs, newest programs, recommended programs, the most downloaded programs, the most viewed

[43] https://play.google.com/store/apps/details?id=org.catrobat.catroid

programs, and random programs. This gives the visitor an overview of the available programs and informs of new prominent programs. Furthermore, the homepage provides easy navigation to these displayed programs. Additionally, every web page includes the internal search function to search for programs, which are either identified by the entered keywords or by tags.

The *program page* contains features for the interaction with the shared programs. These features are the interactive execution of the developed program, the download of the Pocket Code programs and their compiled APK and Pocket Code files. Additionally, this web page provides an overview of the downloaded and modified programs in the remix graph. This remix graph shows the origin of a modified program and every derived version. The code statistic embedded in the program page informs the users about the used building-blocks and provides a summary of the used elements of the program. The social engagement feature, such as the emoticons or the comment function, encourages the audience to interact with the website. The Pocket Code developers are receiving feedback about their work through these social interactions. The social interactions should encourage developers to engage with Pocket Code to create and publish programs.

The *internal search function* offers visitors the possibility to search through the offered programs by keywords or tags. Developers can tag their developed programs with descriptive terms. As a result, website visitors can find similar programs by these tags. For example, such tags are Game, Experimental, Animation, and so forth.

The *profile page* provides functions to edit the profile information and to manage the uploaded programs. The uploaded programs can be deleted or marked as invisible to hide the program from the internal search and the display on the homepage. Furthermore, the program owner can reset the invisible programs. Consequently, community members are capable of finding these programs again.

The *tutorial* section offers the developers resources to assist them in the program creation. The Google Group[44] is a platform for the exchange of thoughts, ideas, or to get help from experienced developers. The tutorial section contains beginner programs for the introduction of the Pocket Code

[44] https://groups.google.com/forum/m/?fromgroups#!forum/pocketcode

language. Moreover, guides are available that are introducing the Pocket Code building-block language through example programs and step-by-step instructions.

The *media library* contains downloadable content for the usage in the created programs. The available media content consists of images for backgrounds, landscapes, portraits, or sounds for an acoustic program response. The media library should encourage developers to use media content in their Pocket Code programs.

9. Trackable Clickstream Content of the Catrobat Community Website

Chapter 5 and chapter 7 introduced the methods to capture data about a visitor's clickstream. Furthermore, web analytics and their corresponding web metrics to extract data about the website's usage from the gathered clickstream were discussed. The following chapter introduces the trackable content of the Catrobat sharing website's web pages, which were outlined in chapter 8. This chapter proposes the trackable clickstream content of each web page respectively. This includes the homepage, search page, login page, program page, profile page, Pocket Code library, and finally the interactive video content.

9.1. Homepage

The homepage, illustrated in Figure 9.2, is the main entry web page for the website visitors and includes the navigation to the underlying website content [Cam14]. The trackable clickstream content of the Catrobat sharing website can be differentiated in four types, as illustrated in Figure 9.1:

- Page Views
- Internal Links
- External Links
- Interactive Content

The standard page tag tracks every page load, as discussed in section 4.2. As a result, the page view tracking does not need a further code

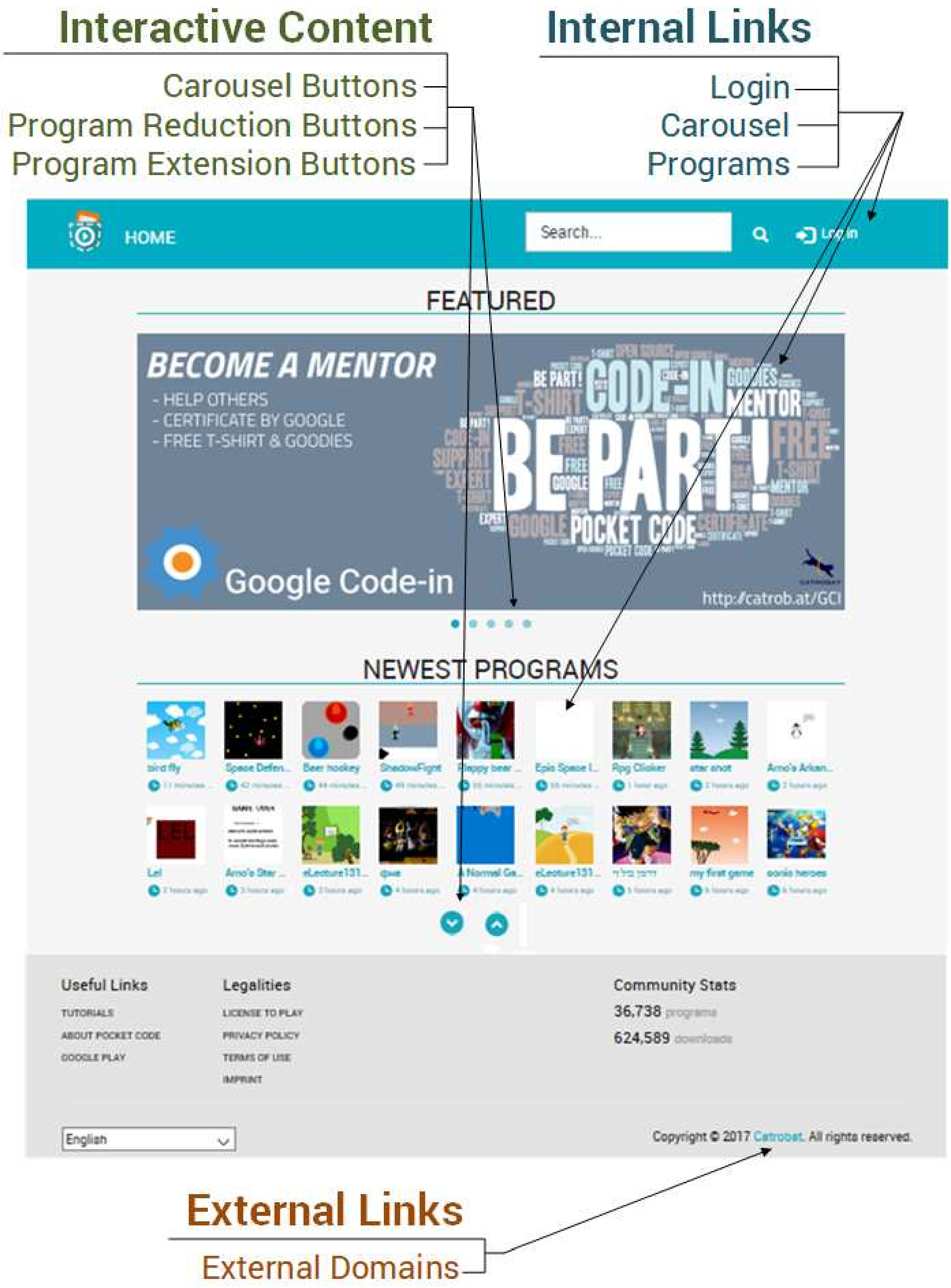


Figure 9.1.: Home Page with different Types of Interactive Content

implementation. Internal and external links, marked in Figure 9.1 in blue and orange respectively, are triggering a full page load. In opposition, the interactive content marked in green only updates a part of the displayed content, and therefore the web browser does not execute a full page load. Furthermore, external links trigger a not trackable page load on the referring website, as discussed in chapter 7. Consequently, interactions with external links must be captured on the referring website before the interaction with the website is terminated. As well as, interactions with the interactive content need a customized tracking code because the standard page tag only tracks full page loads.

Figure 9.2 shows the homepage with potential trackable content in the highlighted areas. Each content of the highlighted areas can be tracked by event tracking. The header of the web page contains the search field marked with the box numbered (1) in Figure 9.2 and is used to search for particular words in the title of the uploaded programs. Additionally, the header contains the login button, which redirects the user to the login page. The main focus of this web page is the presentation of the uploaded programs, grouped by *featured, newest, recommended, most downloaded, most viewed* and *random* program categories marked from (2) to (7) in Figure 9.2. The carousel (2) contains links to Pocket Code programs and other featured information. A click on the carousel (2) redirects the visitor to the targeted web page. The standard page tag only track the visitor's path but it cannot identify the interacted element on the web page. A method to get the information about the interacted element is the event tracking. By tagging each element that contributes to the reconstruction of the visitor path with a custom tracking code, data about the interacted element can be gathered. Subsequently, each interaction with the web page elements, enumerated from (3) to (7), can be identified. Additionally, the buttons with the postfix "a" from (3a) to (7a) in Figure 9.2 are reducing or extending the number of displayed programs of each category. An interaction with these buttons only partially updates the content of the web page. As a consequence, the standard page tag cannot capture these interactions and must be tracked manually by event tracking.

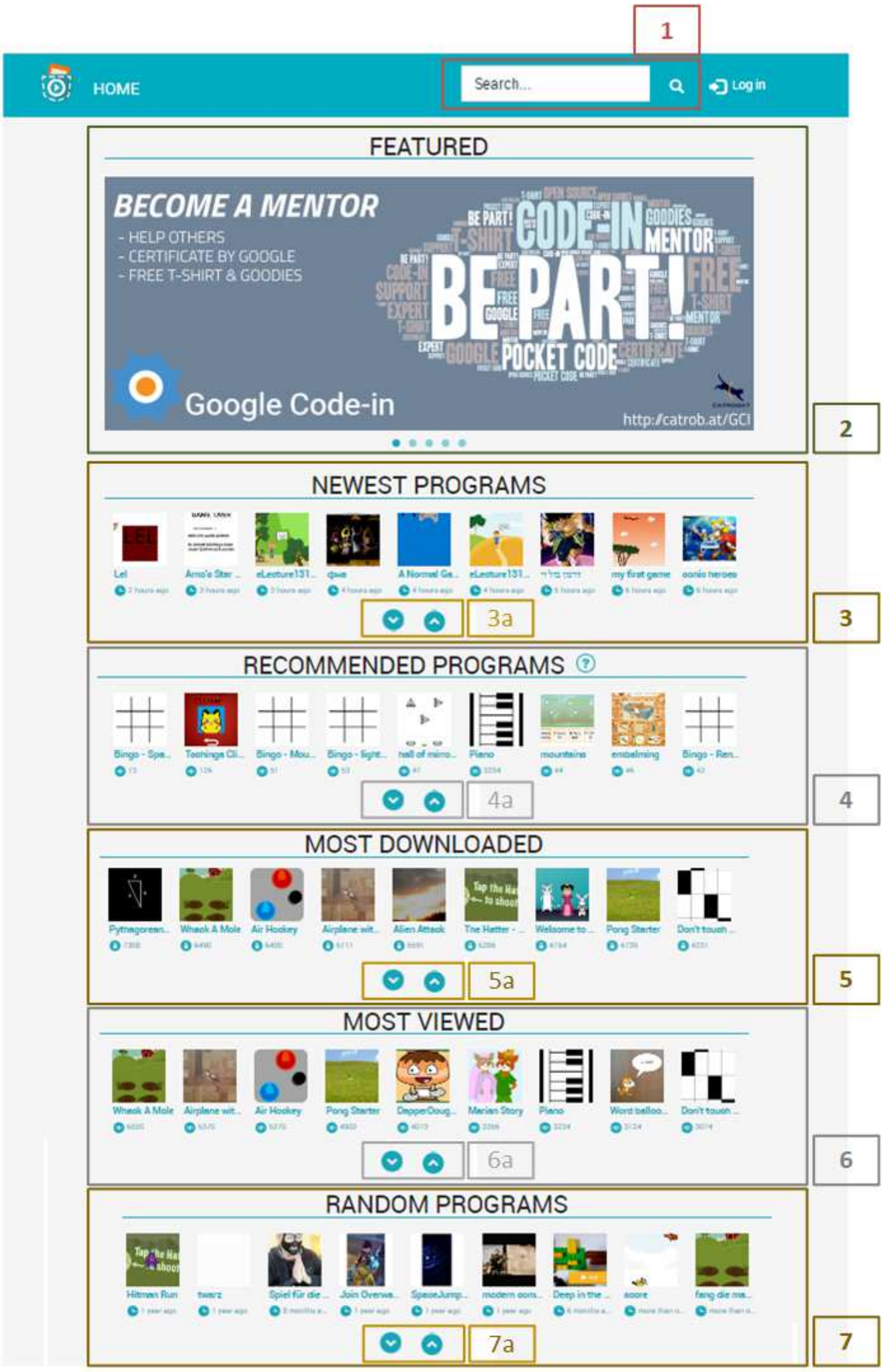
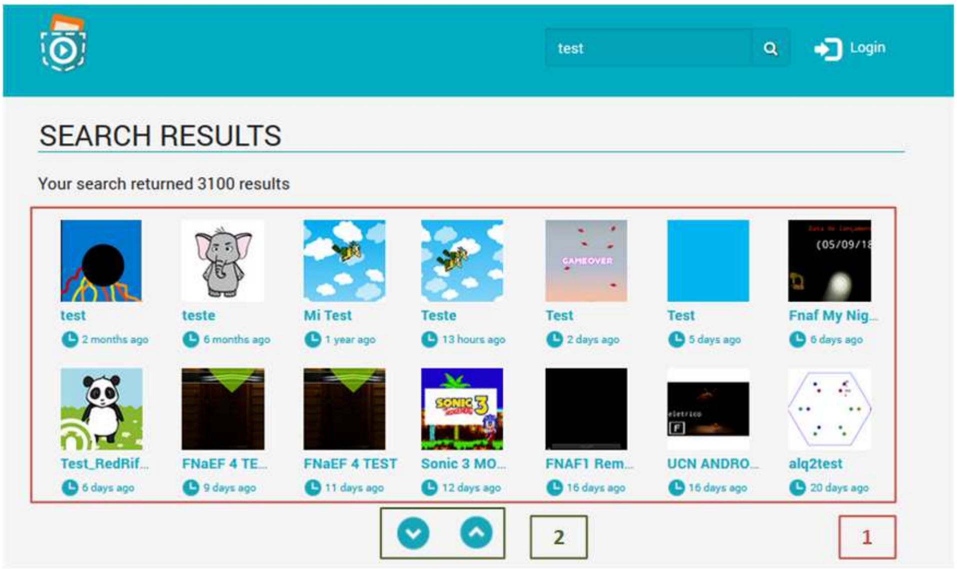


Figure 9.2.: Trackable Content on the Homepage of the Catrobat Community Website

9.2. Search Page



SEARCH RESULTS

Your search returned 3100 results

Figure 9.3.: Trackable Content on the Search Page of the Catrobat Community Website

The search function (1), illustrated in Figure 9.2, is embedded into the header of every web page. This provides visitors the ability to search for programs independent of the viewed web page. The search function redirects the visitor to the search page, illustrated in Figure 9.3, that supports the searching for:

- program names
- tags

The search results are displayed in (1) and displays the search results for the searched keyword. Furthermore, the web page includes the more- or less-button pair (2) to extend or reduce the number of displayed search results and only customized tracking code can track interactions with these buttons. A click on a program in the search results (1) navigates the visitor to the program page for the clicked program. The standard tracking code identifies the navigation from the search page to the program page and the corresponding page path. However, if the individual programs are grouped in a single program group then tracking code must be added to each displayed search result to identify the clicked program.

9.3. Login Page

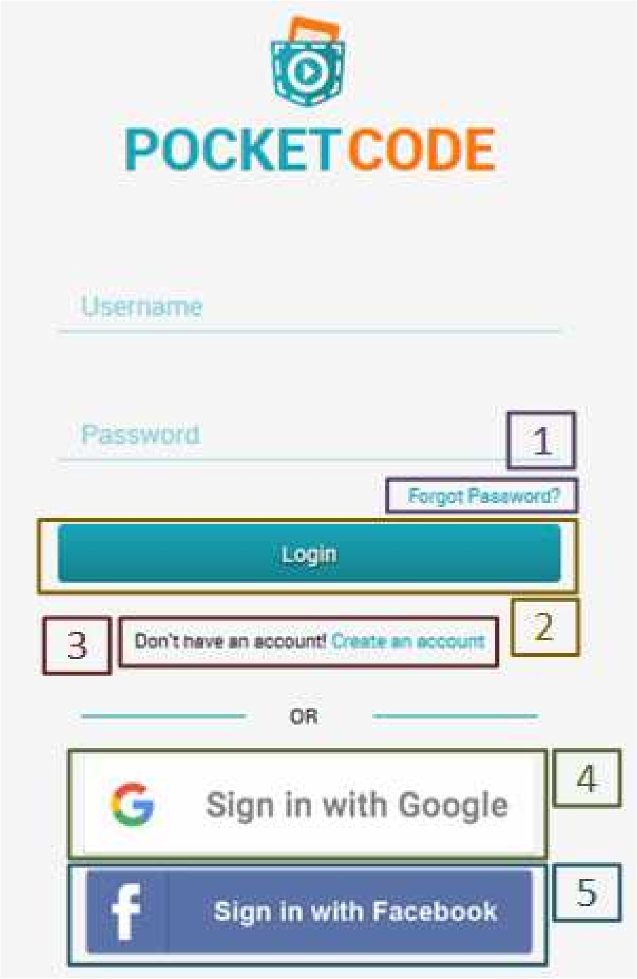


Figure 9.4.: Trackable Content on the Login Page of the Catrobat Community Website

The login page, illustrated in Figure 9.4, provides the register and the log-in function. The supported login methods are via username, Google+ account, or Facebook account. The interactions that can be captured on the login page are marked from (1) to (5). Since a click on *Forgot Password?* (1) and *Create an account* (3) leads to a following unique web page, the user interaction is tracked by the standard page tag. Subsequently, this interaction does not need to be tracked with a custom tracking code. It is necessary to track the interaction with the login button (2), the Google+ button (4), and the Facebook button (5). This ensures that the event metric of the web analytics tool contains the used login method. As a result, the audience is segmentable by the login method, and the differences in the user behavior are identifiable.

9.4. Program Page

The program page, shown in Figure 9.5, provides interactions to download Pocket Code programs (3), generate APK files (5) (Android Package Kit[45]) from the Pocket Code program, view the Remix graph (4), display the code statistics (7), and display the code of the program (8). Furthermore, the integrated HTML5 viewer supports the interactive execution of the program directly in the web browser (1). The code statistic (7) displays an overview of the used code blocks in the Pocket Code program grouped by the category and the Pocket Code building-block composition (8). The associated tags are listed below the description on the web page (2). These tags can be used to search for similar programs with the same tag. Identically to the homepage, the program page provides the categories *similar programs* (9) and *recommended programs* (10) for the simple navigation to relevant programs. Signed-in users can rate uploaded programs via emoticons (6) and comment (11). These interaction are data source to measure the number of social interactions and to compute the social engagement score, which is described in chapter 7.3.

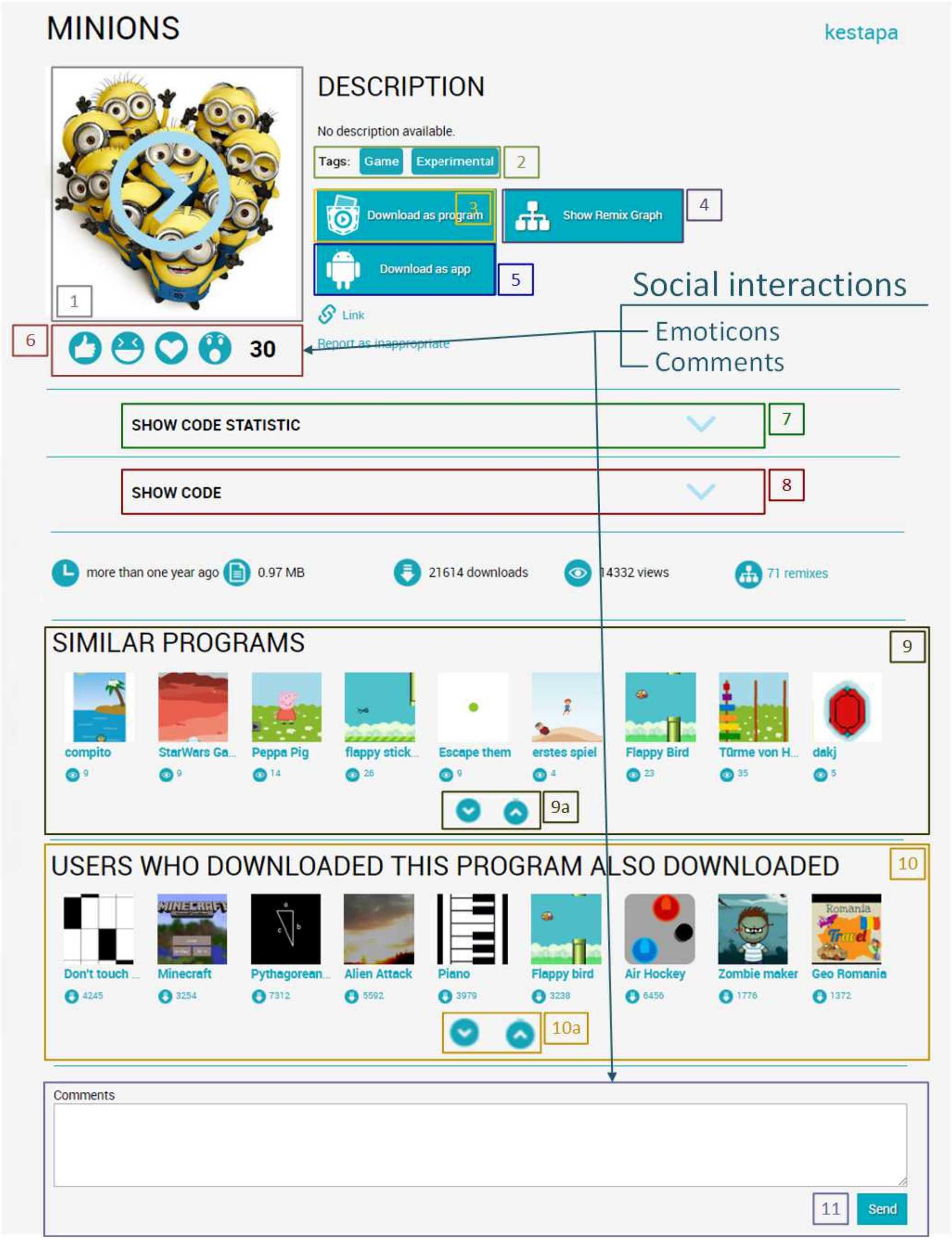


Figure 9.5.: Trackable Content on the Program Page of the Catrobat Community Website

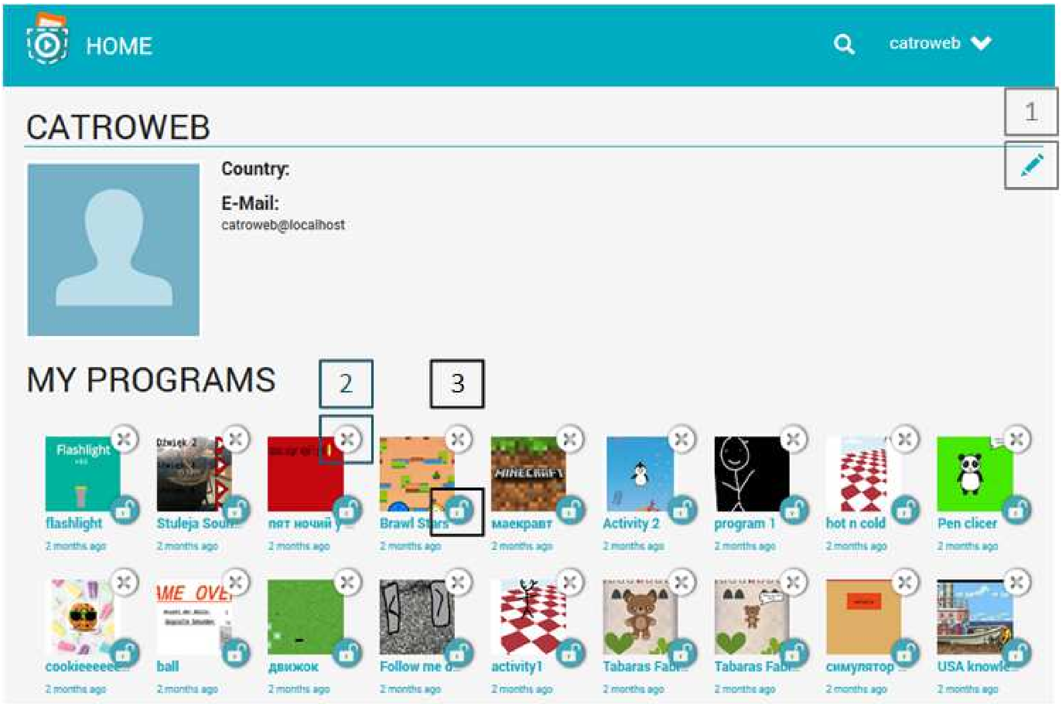


Figure 9.6.: Trackable Content on the Profile Page of the Catrobat Community Website

9.5. Profile Page

The profile page, illustrated in Figure 9.6, serves the purpose to edit the user profile and to manage the uploaded programs. The edit profile button (1) redirects to the edit web page. This behavior is tracked by the standard page tag that monitors the navigation behavior of website users. A click on the cross-icon (2), on the top right corner of each displayed program, deletes the program from the uploaded programs. This action is only trackable by event tracking because no navigation to another web page takes place. Identically to the deletion button, an interaction with the visibility button in shape of a lock-icon (3), on the bottom right corner of each program, is only trackable by event tagging.

9.6. Pocket Library

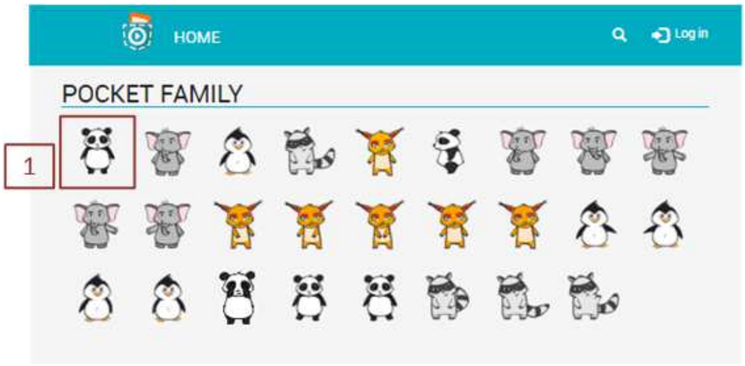


Figure 9.7.: Trackable Content in the Media Library of the Catrobat Community Website

The Pocket Code library provides many different kinds of resources for the creative use in the creation of Pocket Code programs:

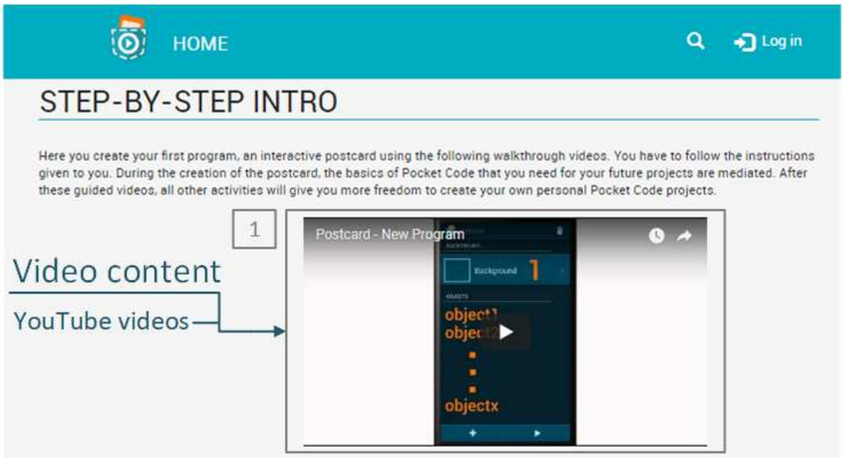
- Looks
- Sounds
- Backgrounds

[45] https://en.wikipedia.org/wiki/Android_application_package

- Landscapes/Portraits

Figure 9.7 illustrates the looks section of the Pocket Code library. Each downloadable resource marked with (1) triggers a JavaScript code that redirects the visitors to the resource. The download of the resource does not trigger a web page load that is trackable by the standard tracking code. However, an implementation of a custom tracking code can be used to track these downloads. This is done by adding a customized JavaScript EventHandler to the click event of every resource to track the download triggered by the JavaScript functions. Information about downloads can then be integrated into the user behavior flow to determine the clickstream of individual website visitors, as discussed in section 6.9.

9.7. Video Content



STEP-BY-STEP INTRO

Here you create your first program, an interactive postcard using the following walkthrough videos. You have to follow the instructions given to you. During the creation of the postcard, the basics of Pocket Code that you need for your future projects are mediated. After these guided videos, all other activities will give you more freedom to create your own personal Pocket Code projects.

Figure 9.8.: Trackable Content of the Step-by-Step Intro of the Catrobat Community Website

Figure 9.8 shows an example web page that contains the video content. On this web page, a step-by-step tutorial video (1) is provided for the building of a simple interactive postcard application. Each video of this web page describes a single feature of the Pocket Code program creation. By adding an EventHandler to specific events, for example, play, pause, buffering, or the video progress (25%, 50%, 100%), actionable interaction data can be collected, as described in section 6.15. The HTML5 video player API provides events, where these EventHandlers can be registered. The user behavior of the video viewers can then be analyzed by measuring the performance of the entire video content or individual videos of the website.

10. Tracked Clickstream Content of the Pocket Code Sharing Website

This chapter discusses the current state of the Pocket Code sharing website's web analytics implementation and illustrates the possibilities of web analytic tools on the use-case of the Pocket Code sharing website. Furthermore, the challenges of the tracking implementation of web analytics are discussed.

10.1. Current State

Currently, the Catrobat sharing website uses JavaScript page tagging, as discussed in section 4.2, to gather data about the audience. The JavaScript page tag tracks the website interaction from the client's perspective. However, only the standard page tracking code is implemented on each web page. The standard tracking code supports the tracking of each web page load to reconstruct the visitor path. This includes metrics, such as the page views, the landing page, the exit page, the behavior flow, sessions, visitors, unique visitors, bounce rate, or depth of visit. Technical information, such as the browser type, the browser version, and the browser language are extractable from the User-Agent of the HTTP-Header. Additionally, the HTTP-Header includes the Referer field (misspelling of referrer[46]) that contains the URL of the last visited web page. This referrer provides information about the type of the hyperlink (internal or external hyperlink), as well as, information about referring websites. Thus, this data is used to identify the traffic

[46] https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Referer

sources of the website. However, data from the page view tracking is currently not used to analyze the visitors or user interactions with the Catrobat community website.

10.2. Possibilities

Web Analytics offers many different vectors to analyze the performance of the website, the users, and the user-interactions. The first vector to be examined is the traffic source. The traffic source metrics, discussed in section 5.3.1, determine the origin of the website visitors by identifying the direct traffic, the referring websites, the social networks, the search engines, or the used search keywords. The determination of the website visitor's traffic source serves the purpose to measure their performances. Moreover, the identification of the traffic sources provides data for the traffic source rating, optimization, or the segmentation of the audience to identify differences in the user behavior. Additional information gather by web analytic tools (for example, geolocation, new visitor, returning visitor, or page depth) provides information about the characteristics of the visitor, as discussed in section 5.3.1. The identification of traffic from unique visitor metric is an important part of the visitor analysis and subsequently, the determination of the ratio of new visitors to returning visitors. The visitor metrics and reports are used to measure the visitor loyalty of the website audience, which can be used to determine the number of visitor sessions, and the days elapsed since the last visit. Segmenting the audience serves the purpose to analyze the behavior of different visitor groups. Further information about the technical device used to interact with the website, the used browser, or browser version is available for technical statistics, as well as, for the segmentation of the audience. Since the Pocket Code application provides its own browser, named *Catrobat*, the identification of the traffic from the Pocket Code application is effortless. Based on the IP-address, the geographic location of the website visitor can be estimated, which provides further data for the segmentation of the audience.

A further investigation of the subsequent web page views offers information about the behavior flow and the user behavior of the website. The interaction behavior of the website visitors can be determined with the metrics and reports categorized by the behavior analysis, as discussed in section 5.3.1. An important indicator for the visitor engagement is the interaction time with individual web pages and the entire web page. This information can be used to identify the web page with high engagement value and to determine web pages, where visitors are experiencing confusions. The page depth metrics provides information on how many web pages a visitor views during a single session. This metric can be used to segment the audience to identify traffic sources and to observe the user behavior of the different segments. The bounce rate measures the engagement value of the landing pages, the search keywords, or the referring websites. The behavior flow graph describes the subsequent web page views of the visitor and supports the visual representation of a visitor's navigation path. The event tracking, described in chapter 7.2, adds additional information to the navigation path, which can be used to analyze the events on each web page. The tracking of the internal search keywords provides information about the user experience and gives insight into user expectations, such as missing features of the website or obstacle in the visitor path.

The content metrics, discussed in section 5.3.1, are identifying the consumed content of the website. The page view metric identifies the frequently viewed content, and by monitoring the number of visitors for a certain period of time. This supports the creation of custom-tailored content, which is adjusted especially for the user needs. Web analytic tools also support the measurement of the site speed, as discussed in section 6.14. This site speed determines the load time of individual web pages and can be used to measure the performance of conducted site speed metric optimizations. Similar to the website content, the video content can be analyzed to identify visitor trend or to ascertain problems of the visitors through the investigation of the video scenes, as discussed in section 6.15.

The social engagement score, introduced in section 7.3, measure the social interactions on the website by assigning a value to each social interaction. The implementation of event tracking, as discussed in section 7.2, complements the page tagging by tracking the interactive website interactions. For example, the tracking of interaction with interactive website elements, the social engagement score, or the video tracking.

10.3. Challenges

As introduced in section 8.2, the target group of the Catrobat project are children and teenager from the age of 13 to 17 years. Since the target group are underaged children, special care must be placed to protect their privacy. For children did the US government issue the COPPA - Children's Online Privacy Protection Act[47] that regularizes the privacy of children under the age of 13. However, the target group of the Catrobat project is children from the age of 13 and as following COPPA does not apply. In May 2018, the EU did issue the General Data Protection Regulation (GDPR or germ. DSGVO), which regulates privacy protection of personal data. As a result, special care must be placed in the collection and processing of personal data. Whereby it concerns only personal data, personal data in the context of web analytics is not clearly defined. Furthermore, the GDPR only handles personal data and impersonal data or anonymized data are not affected. However, the gathering of the personal data the IP-addresses, the telephone number, or the address are affected. Subsequently, the user consent must be acquired through Opt-in to gather this data [Has12; Cli12]. Web analytics tools are gathering the IP-address of the visitor [Cli12], as a consequence, the IP-address must be anonymized to protect the privacy of the visitors. The European Union (EU) did issue the *ePrivacy directive* (2002/58/EC [48]), or cookie directive, to protect the privacy and personal data of website visitors.

According to *"Article 29 Data Protection Working Party"*, the ePrivacy directive regulates that each cookie that is set on the client and is not strictly required for a working client-server communication, the consent of the user must be acquired beforehand [Par12]. As a consequence, some features of web analytics tools are not used for the enhancement of the data accuracy. For example, an addition of a unique identifier, discussed in section 6.2, to the visitor's session during the login. This additional identifier would identify unique visitors across multiple devices or browsers, as outlined in section 6.2. To further enhance the privacy protection, the connection between the

[47] `https://www.ftc.gov/tips-advice/business-center/privacy-and-security/children%27s-privacy`

[48] `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML`

analytic vendor and the client is transmitted via HTTPS, and no personal information is processed via the web analytics implementation.

Google Analytics offers the possibility to gather demographic information about the visitors by tracking visitors across multiple domains with third-party cookies, such as the *DoubleClick cookie*. Information that can be gathered about the visitor's demographics are [Ana19a]:

- **Age:** 18-24, 25-34, 35-44, 45-54, 55-64, 65+
- **Gender:** Male, Female
- **Affinity Categories:** For example: Technophiles, Sports Fans, and Cooking Enthusiasts
- **In-Market Segments:** Product-purchase interests
- **Other Categories:** Contains specializations of the **Affinity Categories**

However, the demographic information for the target audience, which is from 13 to 17 years, as outlined in section 8.2, is not available for the Catrobat community website because demographics data of visitors under the age of 18 is not gathered.

10.4. Tracked Content

As described in chapter 9, the main types of events that are trackable by page tagging are:

- Page Views
- Event Tracking
  - Internal Links
  - External Links
  - Video Content
  - Interactive Content
  - Social Content

10.4.1. Page Views

The tracking of the web page views is the foundation of the visitor analysis. Page tags (displayed in Listing 4.1) are embedded on each web page to track the subsequent web page requests of a visitor. The *visitor path*, beginning from the landing page to the exit page, is reconstructed from these subsequent web page requests. The page view tracking code also gathers additional information, such as the screen resolution, browser window size, device type, or referrer about the visitors, as described in section 4.2. This supports the analysis of the in section 10.2 introduced concepts. Consequently, all web page views of the Catrobat community website are tracked by including the tracking code on every web page.

10.4.2. Internal Links

Page tagging, as discussed in section 4.2, tracks the navigation behavior from internal links during each web page load. However, only the navigation is tracked and not the element that has triggered the navigation. An event tracking code must be added to each element to track these elements, Nonetheless, this is only useful where *actionable insight* for business decisions, technical information or important information for the reconstruction of the *user journey* can be gathered. Otherwise, the web metrics will contain a large amount of available data but only few insights are discoverable.

In chapter 9, the trackable content of the web pages was introduced. Also discussed were the internal links for the navigation, downloads of programs or APKs files, or the download of media resources. Information gathered from these elements supports the creation of the behavior flow graph, introduced in section 6.9, and illustrates the entire visitor journey through the website. The tracked internal links on the website, as discussed in chapter 9, are providing information about the visitor's clickstream for the analysis of the user journey. Additional technical information can also be gained by analyzing these hyperlinks to identify website functions that are accepted and used by the audience. For example, additional tracked elements are on the home page 9.1, the program page 9.4, and the Pocket Code library 9.6.

10.4.3. External Links

The tracking of the external links provides information about the visitor's navigation behavior. The monitoring of the incoming and outgoing traffic supports the analysis of the website environment to gain information about visitor expectations [Kau10; Has12; KSK12]. As a result, all external links are tracked. This includes:

- Outgoing Hyperlinks
- Emails

The navigation path is displayed in the *Visitor Clickstream and Behavior Flow*, introduced in section 6.9, and illustrates the viewed web pages and the exit pages. In combination with event tracking, the targeted third-party websites are identifiable. As a consequence, the web analytics implementation monitors all external links of the Catrobat community website. As described in chapter 9, the footer, illustrated in Figure 9.1, contains an external link to the Catrobat[49] website. Furthermore, the *Carousel* on the home page sporadically contain hyperlinks to third-party websites but consists mostly of internal links to featured programs.

10.4.4. Interative Content

The interactive content is only trackable via event tracking, as described in chapter 7. As a result, each interaction that provides useful insight or is used to determine the user journey, an event tracking code is added to each element. Such interactive elements are on many web pages, such as the *home page* (section 9.1), *login page* (section 9.3), *program page* (section 9.4), *profile page* (section 9.5) and *video content* (section 9.7) described in each of these sections. The behavior flow can then be used to analyze the user journey and to improve the website for visitor needs. The social engagement score (described in section 7.3) on the *program page* is tracked via event tracking. As described in section 7.3, an EventHandler is added to each social element to capture click events. The web analytics tool can then calculate the social engagement score from this transmitted data. Subsequently, this social

[49] `https://catrobat.org`

engagement score provides information about the social commitment of the audience. The tracking of the video content, described in section 6.15, offers information about the media usage of the website. Web pages with video content are introduced in section 9.7. Custom tracking code and EventHandlers (described in chapter 11) are implemented to track this video content. The information gathered from the video content is used to analyze the video usage of the website, as introduced in section 6.15.

11. Implementation of Google Analytics

This chapter discusses the code implementation of Google Analytics on the Catrobat community website. Within the practical part of this thesis, the page view tracking code and event tracking code is added to the Catrobat community website. The Pocket Code sharing website is reachable via the URL `https://share.catrob.at/pocketcode/`. At first, this chapter gives a brief introduction about the structure of an HTML document. Followed by the load order of HTML elements in the browser and the `async` attribute of JavaScript files. Furthermore, the structure of the implementation is illustrated, and the used modules in the code implementation are described. As well as, this chapter introduces the event tracking code used to track the page views.

11.1. Anatomy of the HTML Document

Usually, the server responds with an HTML document after a web page request from the client. However, the server can also respond with files, JSON-objects, partial HTML, or plain text. In the following chapter, only HTML documents are discussed. An HTML document begins with the `<!DOCTYPE html>` tag. The DOCTYPE is a legacy artifact and not used anymore for the validation of the content. Nonetheless, it must be included on every web page. This DOCTYPE tag declares the rules of the HTML document and has to be present to be certified as a valid HTML document. [Net19c]

```html
<!DOCTYPE html>
<html prefix="og:http://ogp.me/ns#">
    <head>
      /* Metadata */
      <meta http-equiv="Content-Type" content="text/html;
         charset=UTF-8"/>
      <meta name="viewport" content="width=device-width,
         initial-scale=1">
      /* Title of the Web Page */
      <title>Pocket Code Website</title>
      /* Icons */
      <link rel="shortcut icon" href="/images/logo/favicon.png"
         />
      /* CSS Stylesheets */
      <link rel="stylesheet" href="/css/pocketcode/index.css"
         media="screen"/>
      /* JavaScript Files */
      <script src="/compiled/bootstrap/jquery.min.js" />
    </head>
    <body>
        <p>This is a web page</p>
    </body>
</html>
```

Listing 11.1: HTML document

The HTML tag is the root element of every HTML document and encapsulates the HEAD and BODY tag, as illustrated in Listing 11.1. The HEAD tag contains information, descriptions, or content that is not displayed on the web page, such as:

- Metadata
- Title
- Icons
- CSS files
- JavaScript files

The *metadata* contains instructions or descriptions for the web browser. For example, keywords that are used by search engine robots to index the web page, instructions for search bots, charset, language, viewport, caching, author, title, et cetera. The *title tag* defines the web page title, which is displayed in the web browser's title bar. Moreover, the HEAD element includes resources, such as icons or CSS files, identified by the link tag to enrich the design of the web pages. JavaScript files are included on the web page by the script tag and are adding interactive functionality to the website. For example, the dynamically *adding, deleting, or modification* of the web page content, reactions to events or entirely new website functionality and behavior. The script tag not only provides the functionality to include JavaScript files but also to directly embed JavaScript code. The body tag contains the displayed website content, such as text, images, buttons, hyperlinks, or video content.

11.2. Load Order

```html
<!DOCTYPE html>
<html prefix="og:http://ogp.me/ns#">
    <head>
      <title>Pocket Code Website</title>
      <script src="/compiled/js-1.js" />
      <script src="/compiled/js-2.js" />
    </head>
    <body>
    </body>
</html>
```

Listing 11.2: Load Order Example

The load order of the HTML document and JavaScript files is demonstrated using the code shown in Listing 11.2 and displayed in Figure 11.1. The browser processes the HTML tags in the order they are encountered on the web page. As a result, the first encountered HTML element gets process first, followed by the head tag, the title tag, and the script tag. As soon as the browser encounters the script tag, the browser stops the parsing of the web page until the JavaScript file *"js-1.js"* is downloaded and executed, as shown in Figure 11.1. After the JavaScript file execution, the browser continuous

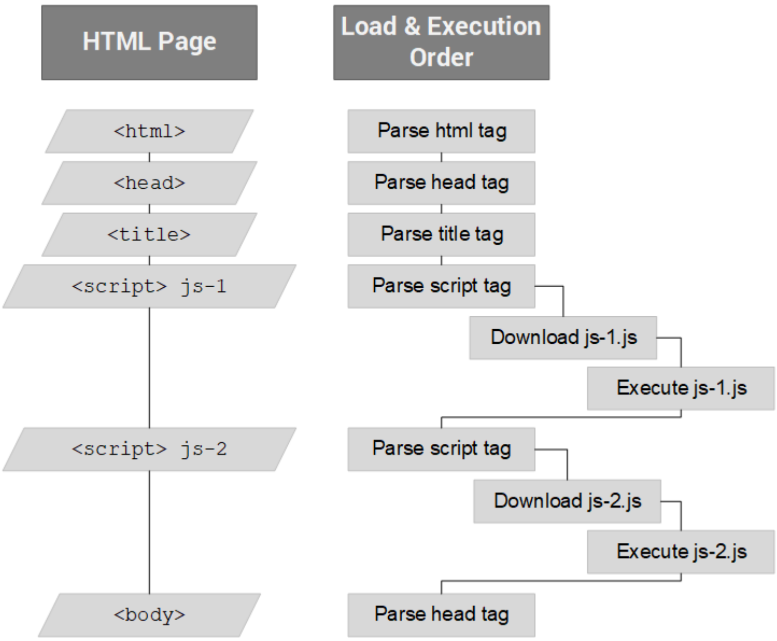


Figure 11.1.: Simplified Load Order of JavaScript Files

the parsing of the HTML document. This parsing suspension occurs for all JavaScript files, such as the *"js-2.js"*. As a result, the loading and rendering of the web page is delayed due to this continuous parsing suspension. The longer loading time has an impact on the user behavior of the visitors and should be avoided, as discussed in section 6.14. Web browsers are providing the *async* attribute for the script tag to prevent this parsing interruption. This *async* attribute instructs the browser to load and execute the JavaScript file asynchronously. However, the *async* attribute is only valid for external JavaScript files and not for inline JavaScript code. [Net19b]

The asynchronous load order of the HTML document is demonstrated using the example code shown in Listing 11.3 and illustrated in Figure 11.2. As displayed in Figure 11.2, the browser does not interrupt the parsing of the HTML document. The web browser downloads and executes the JavaScript files in separate threads, an interruption of the *main thread* does not happen [Net18]. Therefore, the main thread is not blocked by the

```html
<!DOCTYPE html>
<html prefix="og:http://ogp.me/ns#">
    <head>
      <title>Pocket Code Website</title>
      <script async src="/compiled/js-1.js" />
      <script async src="/compiled/js-2.js" />
    </head>
    <body>
    </body>
</html>
```

Listing 11.3: Load Order Async Example

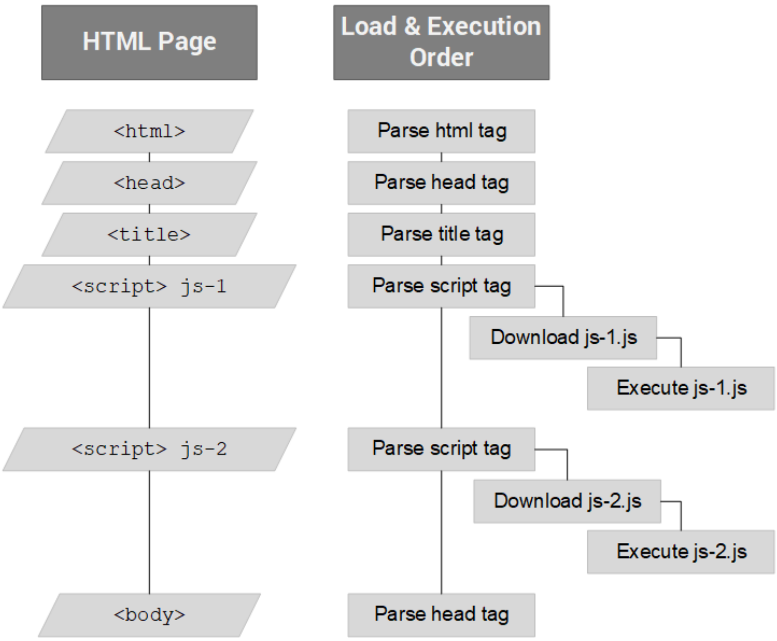


Figure 11.2.: Simplified Asynchronous Load Order of JavaScript Files

JavaScript execution. The web browser executes the JavaScript files as soon as they are downloaded. As a result, the execution order of the JavaScript files is not determined by the order defined in the HTML document. Subsequently, required JavaScript files do not necessarily get executed first even if they are defined first. This could cause dependency errors during the JavaScript execution because required JavaScript dependencies were not loaded. If the load order of JavaScript files must be ensured, the async attribute cannot be used, or a script loader such as RequireJS[50] or yepnope.js[51] must be used for the asynchronous loading of JavaScript files. [Net19b]

11.3. Structure

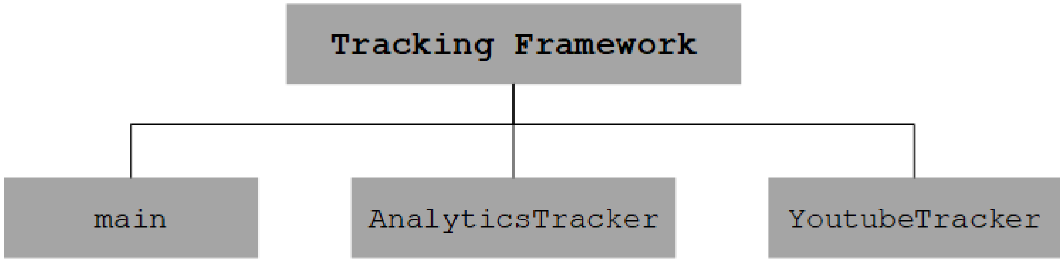


Figure 11.3.: Main Modules of the Tracking Framework

Figure 11.3 illustrates the modules of the implemented tracking framework. The **main** module gets loaded as soon as the analytic script is executed, displayed in Listing 11.4 line 3. The main function is implemented as an Immediately Invoked Function Expression[52] (IIFE) and gets executed immediately after the JavaScript file is downloaded. An advantage of IIFEs is that defined variables in the function scope do not override existing variables with the same name in the global scope. Furthermore, variables defined in the function scope are not accessible from the global scope. This ensures the modularity of IIFE implementations. The main module initializes the Google Analytics Tracking API, the page view tracking, the event tracking, and video tracking, as described in section 11.4. [Net19a]

The **AnalyticsTracker** module offers methods for the registration of JSON-objects (further discussed in section 11.5) that are tracked and adds click-EventHandlers to these HTML elements. Furthermore, if a parameter in the JSON-object is a function, the AnalyticsTracker interprets this function and

[50] https://requirejs.org/

[51] http://yepnopejs.com/

[52] https://developer.mozilla.org/en-US/docs/Glossary/IIFE

returns the result on execution time. The AnalyticsTracker also implements functions for the tracking of outbound links, downloads, and emails.

The **YoutubeTracker** module provides methods for the tracking of YouTube videos. The module implements methods for the communication with the YouTube player API to react to the `onStateChange` event to track interactions with the YouTube video player. The YoutubeTracker observes the state changes of each embedded video and tracks predefined events.

11.4. Page Tracking Code

This section describes the implementation of the tracking code for the Catrobat community website. At first, an example of the tracking code is given followed by a description of the individual parameters of the tracking code for HTML elements.

The JavaScript code in Listing 11.4 shows code samples of the tracking implementation. Line 3 contains the asynchronous load of the analytic script that includes the three modules and is included on every web page of the Pocket Code community website. The file contains the tracking implementation to tracks page views, events, and videos. Line 7 contains the definition of the Google Analytics tracking function **gtag**. This gtag function is used to interact with the Google Analytics tracking API that is loaded on line 10. The `getScript` function of jQuery[53] loads the Google Analytics tracking script asynchronously. The asynchronously loading does not block the main thread, as discussed in section 11.2. As a result, the parsing of the web page content is not interrupted. The second parameter of getScript is a callback function which gets called after the dynamically loaded script is downloaded and executed. This callback executes page tracking code to track every the page load. Further parameter of the page tag are:

- **anonymize_ip**
  The anonymize_ip parameter sets the last 8-bits of the IP-address that

[53] `https://jquery.com/`

```
/* Asynchronous inclusion of the tracking script */
<head>
    <script async src="/compiled/min/analytics.js" />
</head>

/* Definition of the Google Analytics tracking function */
function gtag() { dataLayer.push(arguments); }

/* Dynamic loading of the Google Analytics API */
jQuery.getScript(’https://www.googletagmanager.com/gtag/js?id=UA-42270417-5’, function() {
    gtag(’js’, new Date());

    var s_url =  UrlMapper.map(window.location.href, UrlMapping);
    gtag(’config’, ’UA-42270417-5’, {
        ’anonymize_ip’: true,
        ’page_location’: s_url,
        });
    }
});
```

Listing 11.4: Inclusion of the tracking code

is sent to the data aggregator to zero (0). For example, the IP-address `123.145.223.122` will get changed to `123.145.223.0`

- **page_location**
  The page location defines the URL of the viewed web page, for example, `https://share.catrob.at//pocketcode/program/44132`. Each subsequent event that is sent to the data aggregator after this page view gets assigned to this page_location. Furthermore, the page location is sanitized by the **UrlMapper** thought predefined rules to exclude URL and route parameters. Since Google Analytics creates for each unique page location a corresponding entry in the page views metrics, this is used to group the program pages in a single group to reduce the complexity of the visitor path analysis. For example, the page locations of `.../pocketcode/?q=1` and `.../pocketcode/?q=2` are reduced to one page group `.../pocketcode/`.

11.5. Event Tracking Object

The event tracking object is a JSON-object, illustrated in Listing 11.5. The implementation uses this JSON-object to register the click event handler on the defined HTML element. The event tracking object can have the following parameters:

- BaseSelector
- SubSelector
- Category
- Action
- Label

```
/* Event Tracking Object */
var eventTrackingObject = {
    'BaseSelector' : '#mostDownloaded',
    'SubSelector' : 'a',
    'Category' : 'engagement',
    'Action' : 'clickEvent',
    'Label': function (e) {
        return HelperFunctions
            .removeDomainAndQuery($(e).attr('href'));
    }
}
```

Listing 11.5: Event Tracking Object

This list contains the usually used parameters for the tracking object. However, further parameters, such as *method, value, search_term, transport_type, pathname, file_path* are available. Mandatory parameters are the *BaseSelector*, *Category* and *Action* parameter. The BaseSelector parameter must consist of a jQuery selector[54] to track the HTML element. If the tracked HTML element is dynamical loaded then the root element in the web page must be defined in the BaseSelector and the SubSelector, as a child of the BaseSelector, must be used to identify the dynamically loaded HTML element. The

[54] https://api.jquery.com/category/selectors/

parameter *Category*, *Action*, and *Label* were discussed in section 7.2 and are only outlined. The *Category* parameter defines the assigned event category. The *Action* parameter identifies the event name and should have a unique value. The *Label* parameter can be used to add additional information to the event action. As in Listing 11.5 shown, the tracking object supports the assignment of function to tracking object parameters. These functions are executing as soon as the event is triggered and are overwriting the method with the function result.

```
var trackingObject = TrackingObjectFactory
                        .createObj(eventTrackingObject);
AnalyticsTracker
    .registerElementsForClickTracking(trackingObject);
```

Listing 11.6: Event Tracking Object Registering

Listing 11.6 displays the generation of TrackingObjects from the JSON-object. These TrackingObjects are then registered by the AnalyticsTracker to add click-EventHandlers to each TrackingObject.

11.6. Testing

All Catrobat projects are developed by the test-driven development (TDD) process, where test cases are written before the code is implemented [JS05]. However, executed events by the tracking framework do not change the content of the web page. This has the disadvantage that reactions of triggered events are not observable by Behat[55] tests that are currently used to test the Catrobat community website. The Chrome Browser plug-in *Google Analytics Debugger*[56] was used to test the sending of the triggered events. This plug-in supports the monitoring of the network traffic from the browser to the web server and can be used to inspect the send events.

[55] http://behat.org/

[56] https://chrome.google.com/webstore/detail/google-analytics-debugger/jnkmfdileelhofjcijamephohjechhna?hl=en

12. Analysis of the User Behavior on the Community Website

In this chapter, the gathered data about the visitors of the Catrobat sharing website is used to analyze the user journey from the landing page to the exit page. The gathered data includes information from the page view tracking and the event tracking. Furthermore, the customer journey, outlined in section 12.1, is used as a reference to analyze the user behavior of the website visitors.

12.1. Customer Journey of the Catrobat Community Website

This section gives an overview about the components of the customer journey which was introduced by Stickdorn and Zehrer [SZ09]. The commonly used terms of the customer journey are the touch points, the channels, and the stages. These terms are briefly outlined and the three stages of a website visit: the previsit stage, the visit stage, and the postvisit stage are introduced.

Touch points Touch points are all possible points of interaction that a website visitor can use to directly or indirectly interact with the website [BA17; FKH15]. This includes the referrers, the website navigation, the interactive content, the video content, the social engagement, the login method, uploads, downloads, et cetera.

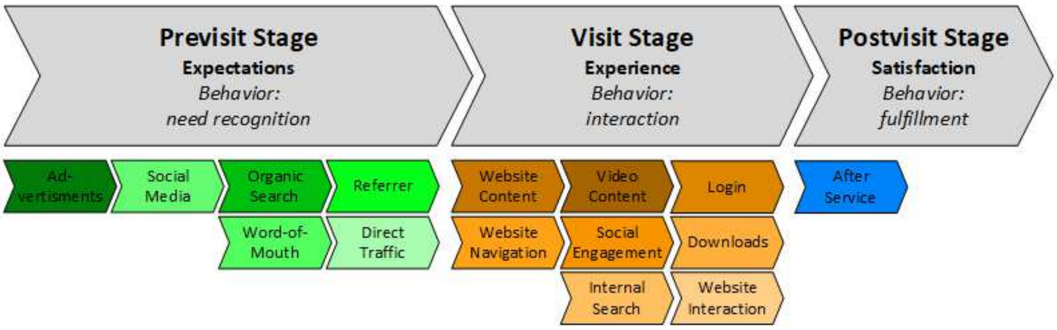


Figure 12.1.: Customer Journey Stages - Catrobat Community Website. Based on [LV16; FKH15; SZ09]

Channels The channel describes the medium chosen to interact with the touch points [BA17; PH06]. In the context of websites, these are the devices used to interact with the website, such as desktops, smartphones, or tablets.

Stages Figure 12.1 illustrates the customer journey stages and identifies the touch point categories of the Catrobat community website. Furthermore, the three stages are representing the entire user experience. Beginning with the identification of the visitor expectations to the exit pages and the reasons for their departure. Moreover, the customer journey stages are monitoring the interactions with Catrobat products aside from the community website. Based on Stickdorn and Zehrer [SZ09] and Lemon and Verhoef [LV16], the touch points are divided into three stages:

- previsit stage
- visit stage
- postvisit stage

The *previsit stage* describes the indirect interactions with the website [SZ09]. These indirect interaction are affecting visitors in their thoughts and are drawing attention to the website [PH06; SZ09]. The web analysis of the traffic sources, as described in in section 5.3.1, tries to data mine these thought from the gathered data and to infer the user expectations. Data sources are the organic and paid search keywords, the content of the referring websites, the percentage of the direct traffic, online marketing campaigns, advertisements, or social media. Further touch points are word-of-mouth [SZ09], and offline advertisements. However, web analytic tools are not able to distinguish this traffic from the occurring direct traffic, and only visitor trends are observable, as described in section 6.16.

The *visit stage* contains all touch points with the website itself [LV16]. Every direct interaction with the website during the visit is a touchpoint and offers information about the navigation behavior of the visitors. These touch points are including every web page load, interactions with the website content, the internal search function, the social content, the media content, the video content, or dynamically tracked events.

The *postvisit stage* contains interactions after the departure from the website [LV16; SZ09]. This includes the download of the Pocket Code application in the Google App Store - *Google Play*, interactions with the Pocket Code application, visits of other resources related to the Pocket Code website, such as newsletters, news groups, forums, blogs, or the YouTube content. Touch points of the postvisit stage can be at the same time touch points of the previsit stage, which leads to the reentering of the visit stage [LV16; SZ09].

12.1.1. The Touch Points

The mapping of the touch points to their corresponding categories in the customer journey is called customer journey mapping [BA17; ROR17] to provide a *"visual representation of the customer processes, needs, and perceptions" (Temkin et al. [Tem+10])*. This customer journey map for the Pocket Code community website is illustrated in Figure 12.2. Based on the entries in the customer journey map, the website visit is analyzed to understand the customer experience [ROR17; LV16] of the website visitors. Information gained from this analysis serves the purpose to improve the customer experience of the provided services [ROR17; LV16; SZ09]. Every touch point category is a group of different touch points to reduce the complexity of the analysis. For example, the category of social media contains multiple social media networks, such as YouTube, Facebook, Google+, Twitter, or Instagram.

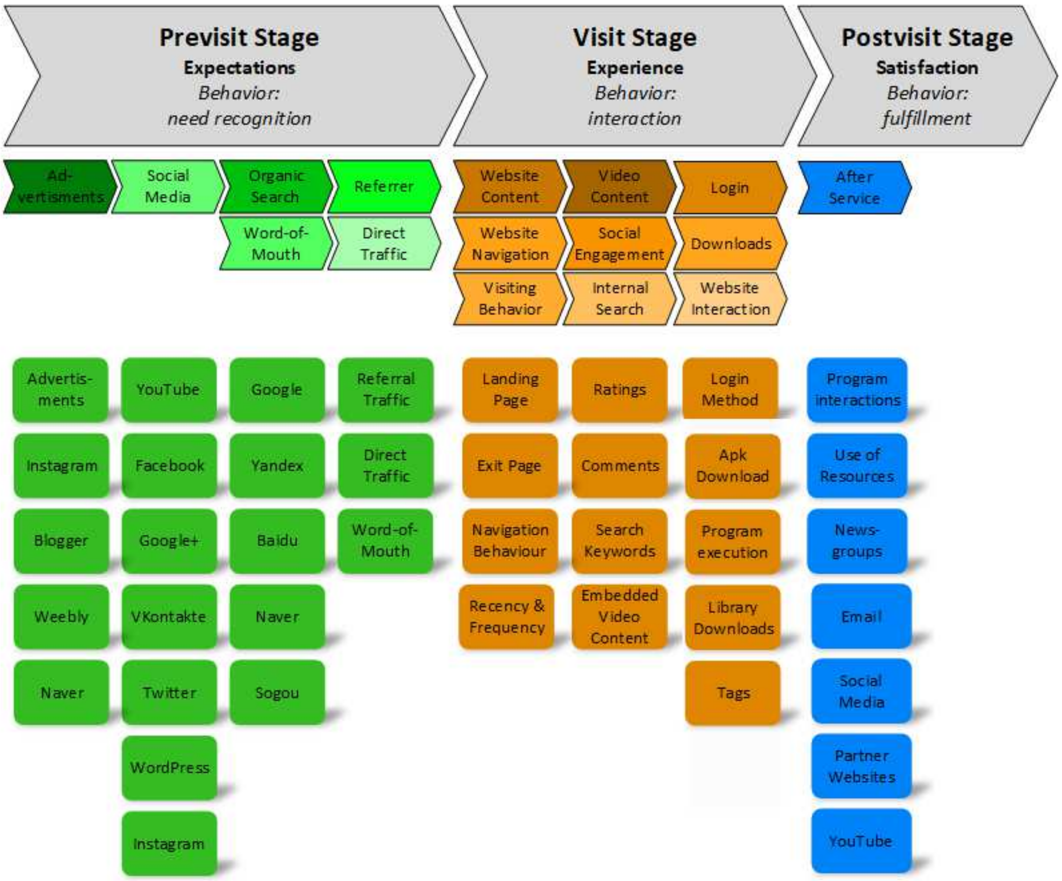


Figure 12.2.: Important Touch Points - Catrobat Community Website. Extended from Figure 12.1 and based on [LV16; FKH15; SZ09]

Based on the information gained from the web analytics tool, 15 touch point categories and more than 40 important touch points are identified, as illustrated in Figure 12.2. The website visitor can visit these touch points represent in the customer journey map zero to multiple times during a website session, even no interaction with entire categories is possible [BA17]. Furthermore, each individual customer journey can have different combinations of touch points [BA17; LV16].

Six touch point categories are identifiable in the previsit stage. This includes traffic sources from *advertisements, social media, organic search, referral traffic, direct traffic, and word-of-mouth*. However, web analytics tools cannot identify the traffic obtained through word-of-mouth, therefore this traffic is included in the direct traffic. Touchpoint categories of the visit-stage are *landing page, the exit page, frequency & recency, the navigation behavior, the consumed content, the video content, the social engagement, the internal search, the login method, the downloaded resources, and the website events*. The postvisit stage consists of the *after service* category. This category contains interactions with related services, which take place after the website interaction. This includes *interactions with the Pocket Code application, the use of resources, news groups, email, social media, partner websites, or YouTube videos*.

12.2. General Information

During the observed time, 66,438 sessions took place on the Pocket Code community website, which are originating from 36,809 users. These 36,809 users are consisting of 33,898 (75%) *New Visitors* and 1,911 (25%) *Returning Visitors*. The average sessions per users are 1.8. Within the 66,438 sessions, 352,513 web pages were viewed, this results in an average page depth of 5.31 pages per visit. The average session duration is 06:57 minutes and the average bounce rate 28.38%.

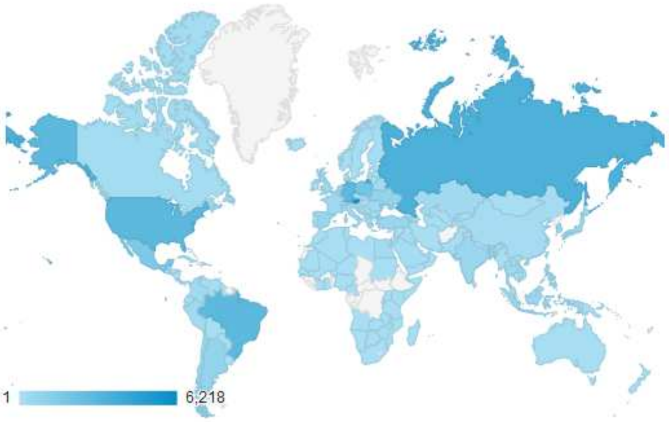
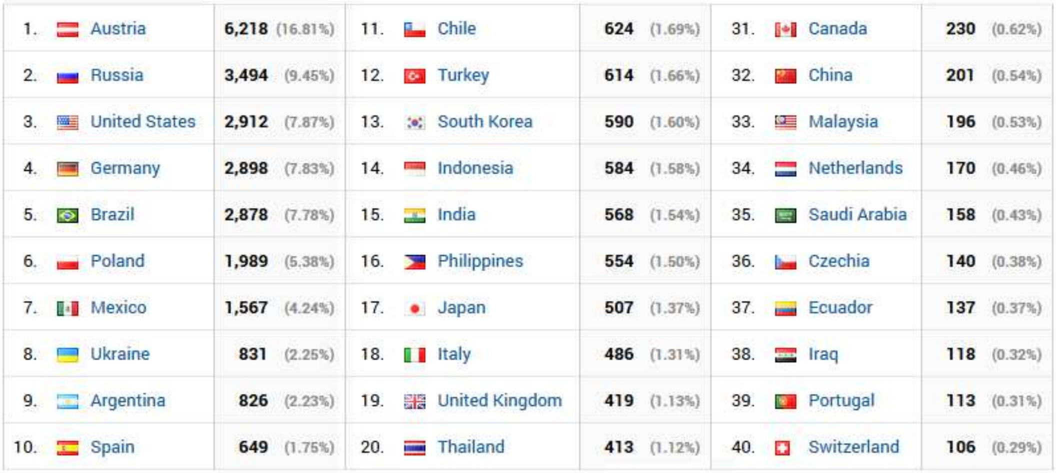


Figure 12.3.: Geographic Map - Catrobat Community Website. Screenshot Google Analytics



1.	Austria	6,218	(16.81%)	11.	Chile	624	(1.69%)	31.	Canada	230	(0.62%)
2.	Russia	3,494	(9.45%)	12.	Turkey	614	(1.66%)	32.	China	201	(0.54%)
3.	United States	2,912	(7.87%)	13.	South Korea	590	(1.60%)	33.	Malaysia	196	(0.53%)
4.	Germany	2,898	(7.83%)	14.	Indonesia	584	(1.58%)	34.	Netherlands	170	(0.46%)
5.	Brazil	2,878	(7.78%)	15.	India	568	(1.54%)	35.	Saudi Arabia	158	(0.43%)
6.	Poland	1,989	(5.38%)	16.	Philippines	554	(1.50%)	36.	Czechia	140	(0.38%)
7.	Mexico	1,567	(4.24%)	17.	Japan	507	(1.37%)	37.	Ecuador	137	(0.37%)
8.	Ukraine	831	(2.25%)	18.	Italy	486	(1.31%)	38.	Iraq	118	(0.32%)
9.	Argentina	826	(2.23%)	19.	United Kingdom	419	(1.13%)	39.	Portugal	113	(0.31%)
10.	Spain	649	(1.75%)	20.	Thailand	413	(1.12%)	40.	Switzerland	106	(0.29%)

Figure 12.4.: Geographic Location - Catrobat Community Website. Screenshot Google Analytics

12.2.1. Geographic Location

As described in section 6.4, web analytics tools are able to determine the geographic location of website visitors. Figure 12.3 displays a world map with the geographic distribution of the website visitors, and Figure 12.4 illustrates the number of unique visitors grouped by the countries. The first two countries in the visitor distribution are Austria and Russia (Google Analytics classifies Russia as a European country) and are from the European continent, followed by the United States of America and Germany. The high amount of sessions from Austria can be explained by the fact that the Catrobat project is mostly developed in Austria. The information gained of the visitor distribution can be used to segment the audience based on the countries. This segmentation supports the analysis of the user behavior of different segments, a comparison of segments, and to target marketing campaigns for individual segments.

12.2.2. Device and Browser

The most used browser to reach the website is the *Catrobat* browser, as illustrated in Figure 12.5. This Catrobat browser is embedded in the Pocket Code application, and used for the interaction between the Pocket Code application and the Catrobat sharing website. Thus, every user of the *Catrobat* browser originates from the Pocket Code application. Following, 56.58% of the users are using the Pocket Code application to interact with the website, as displayed in Figure 12.6. Subsequently, the audience can be segmented in traffic from the Pocket Code application and traffic from web browsers (see Appendix A.3). Additionally, visitors from the Pocket Code application have a low bounce rate of 17.2% compared to the average bounce rate of 28.38% and the bounce rate of web browsers of 48.25%. Furthermore, the direct traffic can be segmented based on the used technology. The Pocket Code application is accountable for 63.90% of the entire sessions. Web browsers, such as Chrome, Safari, Samsung Internet, Android Webview, Firefox, Android Browser, UC Browser, YaBrowser, Edge, Opera, Internet Explorer, and so forth are only accountable for 33.29% of the sessions. Segmented by the direct traffic, the Pocket Code application produces 77.46% of the sessions,

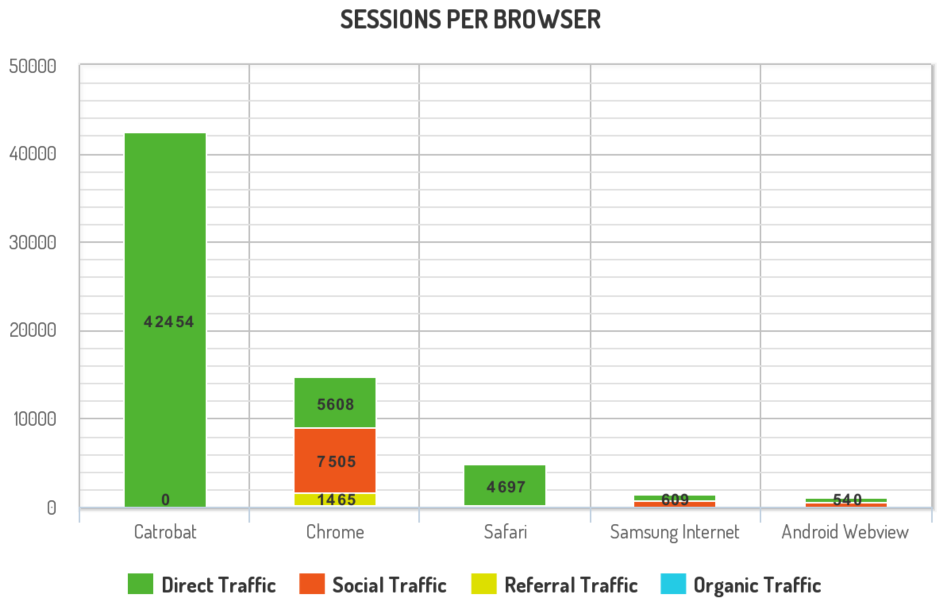
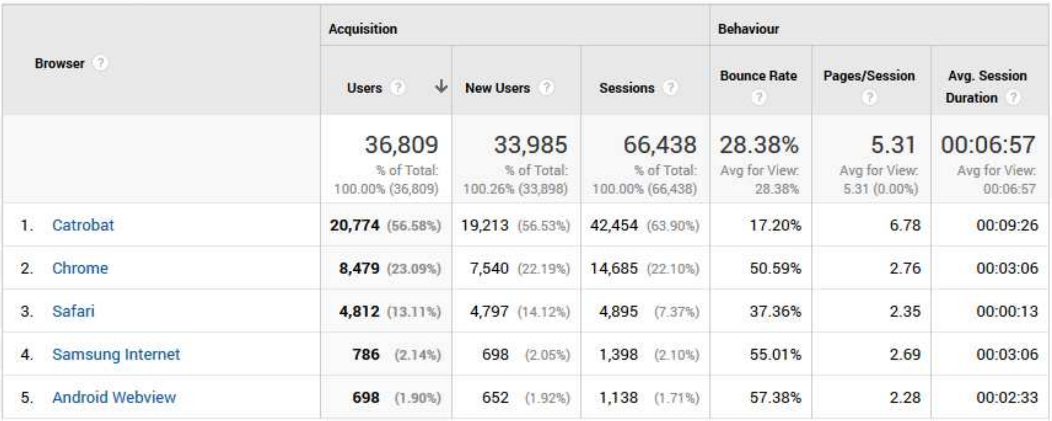


Figure 12.5.: Sessions per Browser - Catrobat Community Website. Source: Google Analytics



Browser	Acquisition			Behaviour		
	Users ↓	New Users	Sessions	Bounce Rate	Pages/Session	Avg. Session Duration
	36,809 % of Total: 100.00% (36,809)	33,985 % of Total: 100.26% (33,898)	66,438 % of Total: 100.00% (66,438)	28.38% Avg for View: 28.38%	5.31 Avg for View: 5.31 (0.00%)	00:06:57 Avg for View: 00:06:57
1. Catrobat	20,774 (56.58%)	19,213 (56.53%)	42,454 (63.90%)	17.20%	6.78	00:09:26
2. Chrome	8,479 (23.09%)	7,540 (22.19%)	14,685 (22.10%)	50.59%	2.76	00:03:06
3. Safari	4,812 (13.11%)	4,797 (14.12%)	4,895 (7.37%)	37.36%	2.35	00:00:13
4. Samsung Internet	786 (2.14%)	698 (2.05%)	1,398 (2.10%)	55.01%	2.69	00:03:06
5. Android Webview	698 (1.90%)	652 (1.92%)	1,138 (1.71%)	57.38%	2.28	00:02:33

Figure 12.6.: Browser Distribution - Catrobat Community Website. Screenshot Google Analytics

and the previously mentioned web browsers are responsible for 22.54% of the sessions, as illustrated in Figure 12.5 and Figure 12.6.

12.3. Previsit Stage

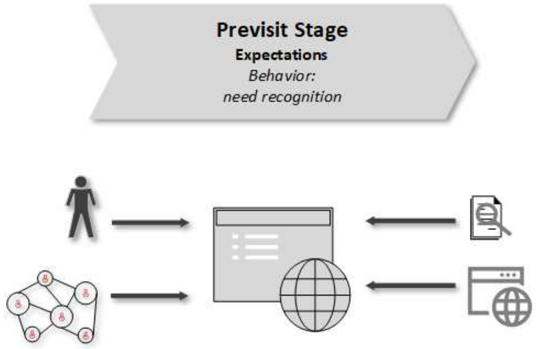


Figure 12.7.: Previsit Stage. Based on [LV16; FKH15; SZ09]

The previsit stage analyzes the characteristics of the visitor's arrival on the Pocket Code sharing website, as illustrated in Figure 12.7. The traffic sources are analyzed based on the reports and metrics introduced in chapter 5.

12.3.1. Traffic Sources

Figure 12.8, Figure 12.9, and Figure 12.10 are displaying the number of sessions for each individual traffic channel. Furthermore, Figure 12.9 illustrates the sessions distribution grouped by the traffic source. The majority of the traffic originates from direct traffic with 82.5% of the total number of sessions. Followed by social traffic with 14%, then referral traffic with 3% and finally traffic from organic search engines with 0.1%, as shown in Figure 12.8.

A high amount of direct traffic indicates that the content of the website is highly rememberable, as discussed in chapter 6.16. A study from *"SearchEngineWatch"* did state that the usual amount of organic search traffic

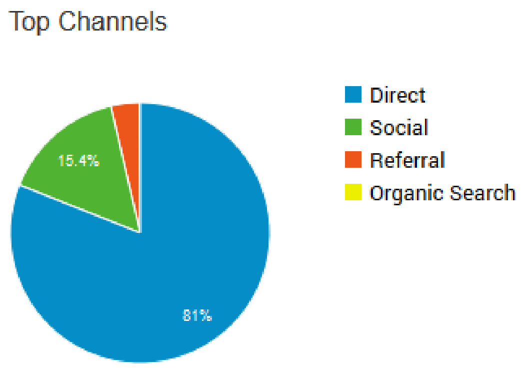


Figure 12.8.: Traffic Channel Distribution - Catrobat Community Website. Screenshot Google Analytics

for websites should be approximately 64% of the total amount of website traffic [Zec14], as outlined in section 6.18. However, the percentage of organic traffic from search engines is only 0.1%. Reasons for this low result can be that the website is not correctly indexed by the search engines, or the crawling of the website is not allowed, as discussed in section 6.18. An analysis of the search keywords used to reach the website is infeasible because of the low amount of available data. Only general assumptions are possible. A further indication for the high engagement of the direct traffic is the low bounce rate of 24%, followed by the organic search traffic with a bounce rate of 37%. The low bounce rate indicates that the used search keywords are highly affine for the content of the website. The bounce rate of the referral and social traffic source is higher than the average bounce rate and amounts to 43% and 48% respectively.

Traffic that originates from direct traffic feature the highest page view count, with an average of 5.8 page views per session, as illustrated in Figure 12.10. Furthermore, the direct traffic has the highest average session duration with 7:46 minutes, followed by the organic search traffic with approximately half the interaction time of 3:49 minutes. This further supports the assumption that the direct traffic has the highest engagement value because of the low bounce rate, the highest number of pages visited per session, and the longest average session duration. The organic search traffic has the second highest

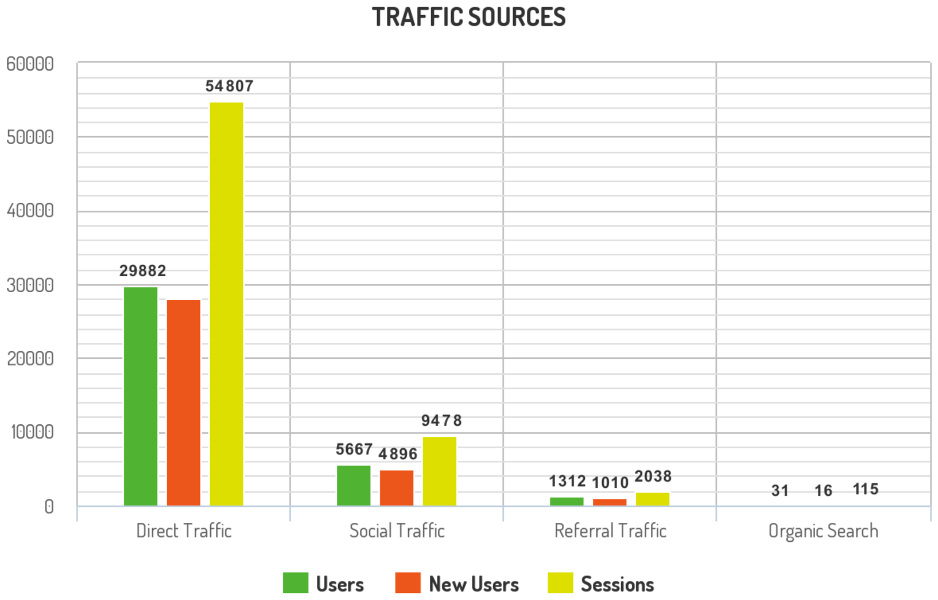
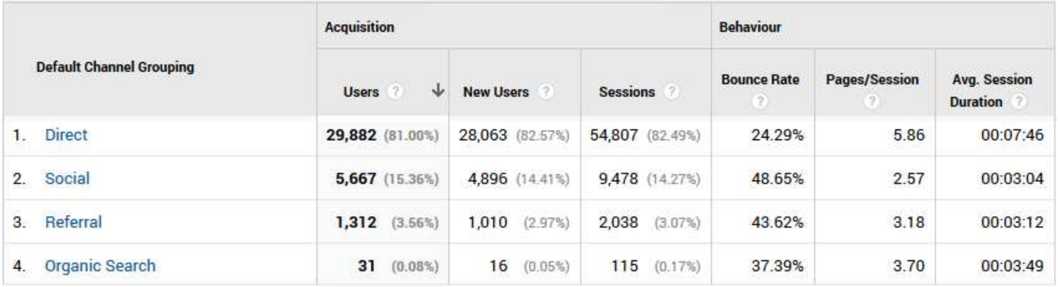


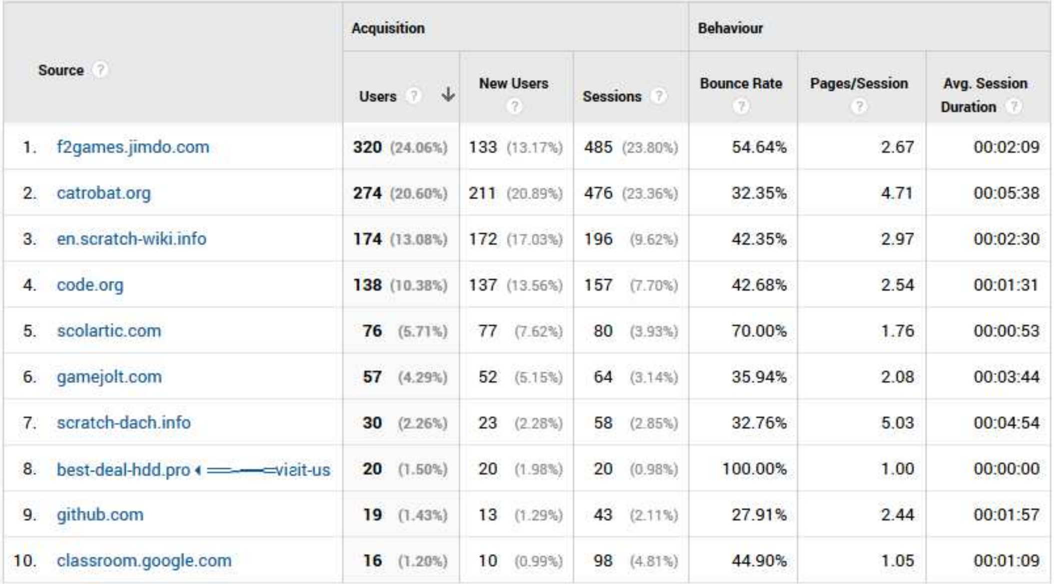
Figure 12.9.: Traffic Channels grouped by Traffic Source - Catrobat Community Website. Source: Google Analytics



Default Channel Grouping	Acquisition			Behaviour		
	Users ? ↓	New Users ?	Sessions ?	Bounce Rate ?	Pages/Session ?	Avg. Session Duration ?
1. Direct	29,882 (81.00%)	28,063 (82.57%)	54,807 (82.49%)	24.29%	5.86	00:07:46
2. Social	5,667 (15.36%)	4,896 (14.41%)	9,478 (14.27%)	48.65%	2.57	00:03:04
3. Referral	1,312 (3.56%)	1,010 (2.97%)	2,038 (3.07%)	43.62%	3.18	00:03:12
4. Organic Search	31 (0.08%)	16 (0.05%)	115 (0.17%)	37.39%	3.70	00:03:49

Figure 12.10.: Traffic Channels Details - Catrobat Community Website. Screenshot Google Analytics

number of page views per sessions.

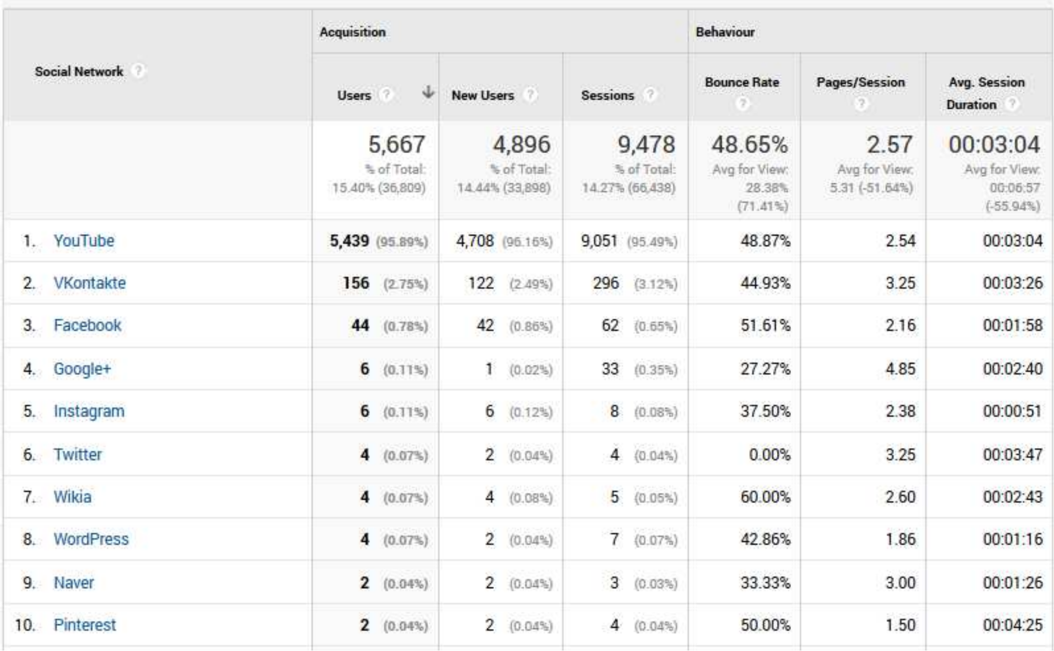


	Source	Acquisition			Behaviour		
		Users	New Users	Sessions	Bounce Rate	Pages/Session	Avg. Session Duration
1.	f2games.jimdo.com	320 (24.06%)	133 (13.17%)	485 (23.80%)	54.64%	2.67	00:02:09
2.	catrobat.org	274 (20.60%)	211 (20.89%)	476 (23.36%)	32.35%	4.71	00:05:38
3.	en.scratch-wiki.info	174 (13.08%)	172 (17.03%)	196 (9.62%)	42.35%	2.97	00:02:30
4.	code.org	138 (10.38%)	137 (13.56%)	157 (7.70%)	42.68%	2.54	00:01:31
5.	scolartic.com	76 (5.71%)	77 (7.62%)	80 (3.93%)	70.00%	1.76	00:00:53
6.	gamejolt.com	57 (4.29%)	52 (5.15%)	64 (3.14%)	35.94%	2.08	00:03:44
7.	scratch-dach.info	30 (2.26%)	23 (2.28%)	58 (2.85%)	32.76%	5.03	00:04:54
8.	best-deal-hdd.pro ◂ ═══visit-us	20 (1.50%)	20 (1.98%)	20 (0.98%)	100.00%	1.00	00:00:00
9.	github.com	19 (1.43%)	13 (1.29%)	43 (2.11%)	27.91%	2.44	00:01:57
10.	classroom.google.com	16 (1.20%)	10 (0.99%)	98 (4.81%)	44.90%	1.05	00:01:09

Figure 12.11.: Referral Traffic Details - Catrobat Community Website. Screenshot Google Analytics

Figure 12.11 shows websites that are referring to the Pocket Code sharing website. Appendix A.2 contains detailed information about the referring websites. The *f2games.jimdo.com* website is accountable for the highest amount of sessions. This website is made by an active Pocket Code developer, who uses this website to publishes the created Pocket Code programs. The *catrobat.org* website generates the second most amount of sessions, and has the longest interaction duration of 06:38 minutes, followed by *scratch-dach.info* with the second longest interaction duration, but the highest amount of page views per session. The worst performance has the website *best-deal-hdd.pro* with a bounce rate of 100%. This indicates that user expectations from the previous website are not met.

An analysis of the used social networks concluded that 95% of the social traffic originates from *YouTube*, as illustrated in Figure 12.12. The second most used social network *VKontakte* only has a traffic share of 3.12%. The traffic source metric hardly represented the the social media network *Facebook* 0.65%, *Google+* 0.35%, *Instagram* 0.08%, or *Twitter* 0.04%. Each of these



Social Network	Acquisition			Behaviour		
	Users	New Users	Sessions	Bounce Rate	Pages/Session	Avg. Session Duration
	5,667 % of Total: 15.40% (36,809)	4,896 % of Total: 14.44% (33,898)	9,478 % of Total: 14.27% (66,438)	48.65% Avg for View: 28.38% (71.41%)	2.57 Avg for View: 5.31 (-51.64%)	00:03:04 Avg for View: 00:06:57 (-55.94%)
1. YouTube	5,439 (95.89%)	4,708 (96.16%)	9,051 (95.49%)	48.87%	2.54	00:03:04
2. VKontakte	156 (2.75%)	122 (2.49%)	296 (3.12%)	44.93%	3.25	00:03:26
3. Facebook	44 (0.78%)	42 (0.86%)	62 (0.65%)	51.61%	2.16	00:01:58
4. Google+	6 (0.11%)	1 (0.02%)	33 (0.35%)	27.27%	4.85	00:02:40
5. Instagram	6 (0.11%)	6 (0.12%)	8 (0.08%)	37.50%	2.38	00:00:51
6. Twitter	4 (0.07%)	2 (0.04%)	4 (0.04%)	0.00%	3.25	00:03:47
7. Wikia	4 (0.07%)	4 (0.08%)	5 (0.05%)	60.00%	2.60	00:02:43
8. WordPress	4 (0.07%)	2 (0.04%)	7 (0.07%)	42.86%	1.86	00:01:16
9. Naver	2 (0.04%)	2 (0.04%)	3 (0.03%)	33.33%	3.00	00:01:26
10. Pinterest	2 (0.04%)	2 (0.04%)	4 (0.04%)	50.00%	1.50	00:04:25

Figure 12.12.: Social Networks Traffic - Catrobat Community Website. Screenshot Google Analytics

social media networks combined, only have a traffic share of less than 1%. The presence of the Catrobat project on these social networks is deficient and should be improved to attract more visitors from these social media networks. The best performing traffic sources are the *direct traffic* (82.49%) and traffic from the video streaming platform *YouTube* (13.11%) (see Appendix A.1 for a detailed statistic). These two traffic sources are responsible for *95.6%* of the *entire traffic* for the Catrobat sharing website.

12.4. Visit Stage

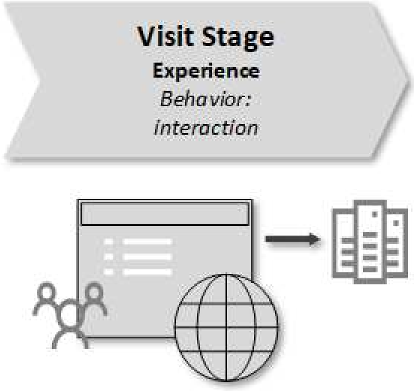


Figure 12.13.: Visit Stage. Based on [LV16; FKH15; SZ09]

The visit stage addresses the behavior of the website visitors during the interaction with the Pocket Code website, as illustrated in Figure 12.13. This includes characteristics of website visitors and of the visit. Furthermore, the frequency and recency report investigates the frequency of visits of returning visitors, and the most frequented landing pages are identified. Afterwards, the user journey of the direct traffic that has the highest engagement value is compared to the general user journey.

Table 12.1.: Average Page Load Time per Browser

Browser	Avg. Page Load Time (sec)
Edge	3.25
Firefox	3.39
Samsung Internet	4.79
Safari	4.92
Catrobat	7.13
Chrome	7.88
YaBrowser	12.92
Opera	14.00
Android Browser	15.51
Android Webview	16.67

12.4.1. Page Load Time

The *average page load time* of the website is *7.15 seconds*. As in section 6.14 outlined, a page load time of more than two to three seconds reduced the revisiting behavior of the audience. Consequently, reducing the page load time should increases the customer satisfaction. Table 12.1 illustrates the average page load time of the different web browsers. The longest average page load time has the Android Webview and the Android Browser with over 15 seconds and the fastest browsers are Edge and Firefox. Web pages with the longest page load time are web pages of the media library, as displayed in Table 12.2. The backgrounds-landscape web page needs for a complete web page load on average 21.03 seconds. The average page load time of the program pages is 9.10 seconds. However, a program page has a page load time of 118.74 seconds, which is the highest single page load of the website.

12.4.2. Frequency and Recency

A significant indicator of the success of websites especially of non-profit's is the visitor loyalty, as discussed in section 6.3. Appendix A.4 and Appendix A.5 are displaying the *Days Since Last Session* and the *Count of Sessions*

Table 12.2.: Average Page Load Time of Web Browsers per Page

Web Page	Avg. Page Load Time (sec)
pocket-library/backgrounds-landscape	21.03
pocket-library/looks	14.15
pocket-library/backgrounds-portrait	9.39
pocket-library/sounds	8.64
program	9.10
help	6.92
pocketcode	6.82
tutorialcards/1	6.70
myprofile	6.45
search	3.09

report. This data supports the segmentation of the audience in the segments displayed in Table 12.3 and Table 12.4 with their corresponding ratios of those segments. The data from Table 12.3 and Table 12.4 of different time periods can be compared to identify visitor trends.

Table 12.3.: Count of Sessions Percentage

Type	Sessions	Percentage
One-time Visitors	1	51.15%
Interested Visitors	2-3	20.99%
Loyal Visitors	4+	27.86%

Table 12.3 segments the audience in three visitor groups the *One-time Visitors*, *Interested Visitors*, and *Loyal Visitors*. *One-time Visitors* are only visiting the website once (one session/visit) and have a share of 51.15% of the total amount of sessions. *Interested Visitors* are visiting the website two to three times (sessions/visits) and have a share of 20.99% of the sessions. 27.86% are visiting the website four or more times and are *Loyal Visitors*. This Table 12.3 shows that most of the visitors 51.15% are visiting the website only once. Causes for this can be that the visitor's satisfaction is not high enough for a further website visit because of the long loading time, website content is not engaging enough, or visitor expectations are not met. Furthermore, the deletion of the persistent cookie to identify returning visitors can also

be responsible for the low percentage of returning visitors, as discussed in section 6.2.

Table 12.4.: Days Since Last Session Percentage

Type	Days Since Last Session	Percentage
Daily Visitors	≤ 1	81.18%
Regular Visitors	1-14	14.66%
Occasional Visitors	15+	4.16%

Table 12.4 segments the visitors in groups based on the days elapsed since a unique user visited the website the last time. 81.18% of the returning visitors, the *Daily Visitors*, are visiting the website on the same day more than once. The time period for the *Regular Visitors* is chosen to be between one and 14 days and have a share of 14.66%. 4.16% *Occasional Visitors* are revisiting the website after 15 days or more. Based on the data from Table 12.4, returning visitors are revisiting the website in a daily to regular fashion. Only a small amount of returning visitors (4.16%) are visiting the website after 15 days.

12.4.3. Landing Pages

Table 12.5 displays the top 10 landing pages segmented by the used browser to reach the website. The Table 12.5 contains the ranking and the names of the landing pages with their corresponding bounce rates. Appendix A.8 contains detailed information about the top 35 landing pages segmented by the traffic from the Pocket Code application and the web browsers. As displayed in Table 12.5, the top landing pages differ based on the technology used to reach the website. For example, visitors from the Pocket Code application arrive at different landing pages compared to visitors from browsers. As a reference, Table 12.5 also contains the popular landing pages of the entire audience. Pocket Code application users mostly arrive at the *homepage*, web pages from the *media library*, and the *program* web page. Users of web browsers arrive on the *program* page, the *help* page, the *homepage*, the *StepByStep* guide, and the *search* web page. The average bounce rate of the

Table 12.5.: Top 10 Landing Pages (App, Browser, All Devices) - Catrobat Community Website

Rank	Pocket Code App	Bounce Rate	Browser	Bounce Rate	All Devices	Bounce Rate
1	pocketcode/	13.46%	program	49.43%	pocketcode/	14.97%
2	pocket-library/looks	18.52%	help	58.65%	program	48.91%
3	pocket-library/sounds	25.25%	pocketcode/	33.86%	pocket-library/looks	18.53%
4	pocket-library/backgrounds-portrait	14.72%	stepByStep	67.44%	help	58.63%
5	pocket-library/backgrounds-landscape	23.11%	search	43.27%	pocket-library/sounds	25.25%
6	program	43.74%	login	62.24%	pocket-library/backgrounds-portrait	14.75%
7	resetting/request	25.23%	termsOfUse	68.13%	pocket-library/backgrounds-landscapes	23.11%
8	search	2.25%	tutorialcards	52.05%	search	22.35%
9	create@school/	11.29%	gaming-tutorials	50.00%	stepByStep	67.44%
10	help	56.67%	pocketgalaxy/	35.59%	resetting/request	25.68%
	Avg. Bounce Rate	23.42%	Avg. Bounce Rate	52.06%	Avg. Bounce Rate	31.96%

top 10 landing pages of the Pocket Code application is 23.42%, the bounce rate of web browsers 52.06%, and the unsegmented bounce rate 31.96%.

12.4.4. User Journey

As discussed in section 3.3, the goal of the behavior analysis (clickstream analysis) is to *infer the intent* of the website visitors. Based on the landing pages, two assumptions about the intents are identifiable:

- to download a resource
- to download a program

In the following section, the behavior flow graph demonstrates the user journey of these two intents. Section 12.4.4 illustrates the user journey to download a program of the unsegmented audience, and section 12.4.4 demonstrates the user journey to download a resource segmented by traffic of the Pocket Code application. For each web page, events are listed and ordered by their frequency. The behavior flow graph segments these events by the device and does not consider the previous user journey to increase the number of covered visitors, as described in section 6.9. Furthermore, the report displays the search keywords for each web page, as discussed in section 6.10.

Download Program - All Users

Figure 12.14 displays the most common user journey through the website with the intent to download a Pocket Code program beginning from the most frequented landing page the homepage. The top most interacted events are the newest programs, recommended programs, newest programs more button, random programs more button, and recommended programs more button. Furthermore, visitors are navigating to programs of interest by interacting with the programs listed in the *newest programs* or *recommended programs* section. Internal search keywords on the landing page are Slendytubbies, Fnaf, Cuphead, Slendytubbies 3, Fnaf 6, Minecraft, Fortnite, Sonic, Undertale, or Mario and each of these keywords can be categorized as a *name of a program*. The homepage's bounce rate of 14.97%, compared to other web pages is low. Only 14.97% of the visitors are terminating the session at this point without a single website interaction.

45% of the visitors are navigating to a program page, and 29% of the visitors are using the search function on the landing page. Visitors on the search page are using the *more, and less buttons* to 13%. Furthermore, 72% visitors of the search web page are using the search results to navigate to a program. 12% of the visitors are revisiting the landing page, and 5.9% are terminating the session.

Most of the interaction takes place on the program page, where 48% of the web page visitors are *downloading* a Pocket Code program. In contrast, only 8.37% are downloading the program as an *APK file*, while 5% are viewing the *remix graph*, 3% are navigating to the next program by clicking on a program grouped by *similar programs*, and 2% by *recommended programs*. The least used elements of the program web page are the *code statistic* with 2.9% and the *interactive execution* of the program in the web browser with 1.7%

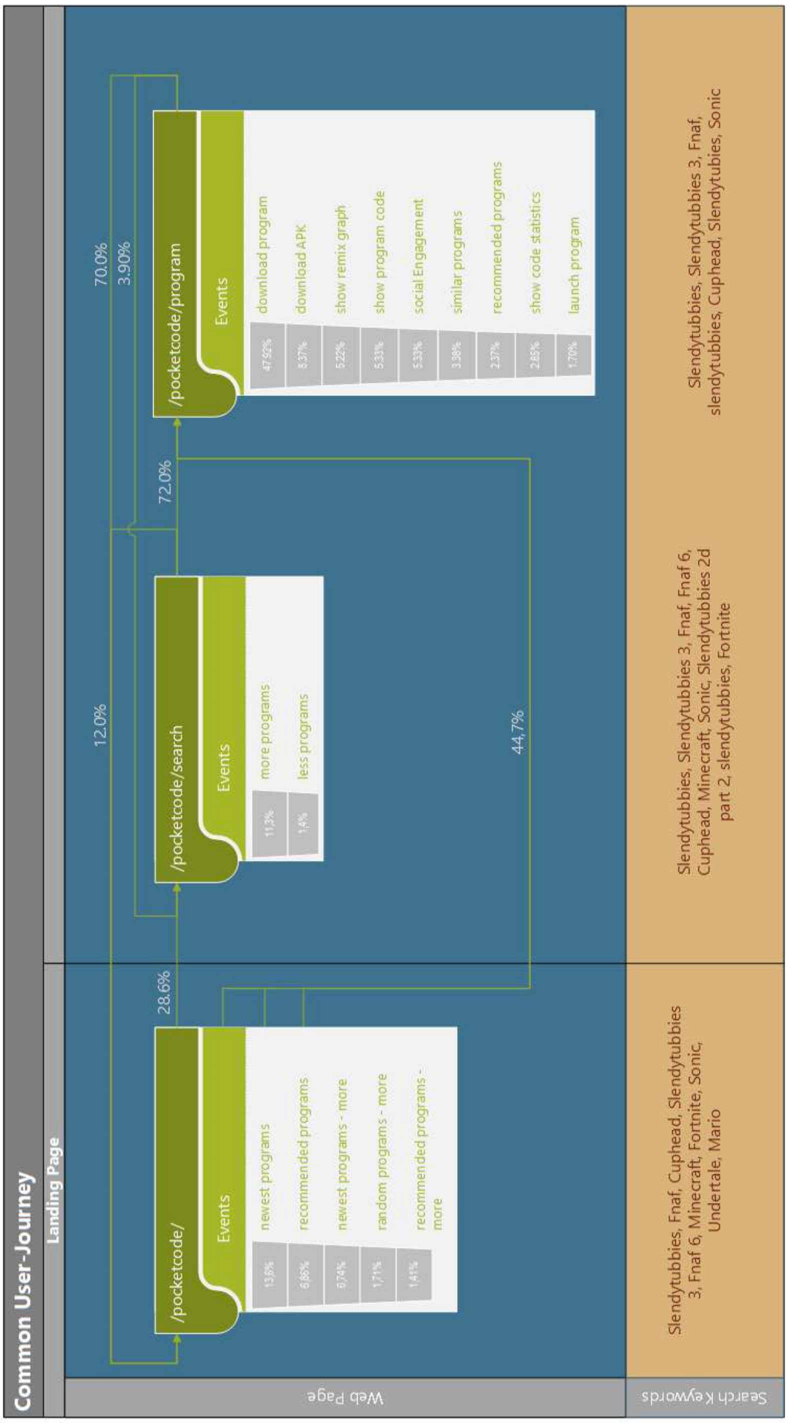

Common User-Journey
Landing Page
Web Page
Search Keywords
/pocketcode/
Events
13,6% newest programs
6,86% recommended programs
6,74% newest programs - more
1,71% random programs - more
1,41% recommended programs - more
28.6%
/pocketcode/search
Events
11,3% more programs
1,4% less programs
12.0%
44,7%
72.0%
/pocketcode/program
Events
47.92% download program
8.37% download APK
5.22% show remix graph
5.33% show program code
5.33% social Engagement
3.38% similar programs
2.37% recommended programs
2.85% show code statistics
1.70% launch program
70.0%
3.90%
Slendytubbies, Fnaf, Cuphead, Slendytubbies 3, Fnaf 6, Minecraft, Fortnite, Sonic, Undertale, Mario
Slendytubbies, Slendytubbies 3, Fnaf, Fnaf 6, Cuphead, Minecraft, Sonic, Slendytubbies 2d part 2, slendytubbies, Fortnite
Slendytubbies, Slendytubbies 3, Fnaf, slendytubbies, Cuphead, Slendytubbies, Sonic

Figure 12.14.: User Journey - All Users - Catrobat Community Website

of the web page visitors. Additionally, 5.33% of the web page visitors are interacting with the *social content*.

After the interaction with the program page, 70% of the visitors are returning to the landing page, 4% to the search web page and 14.4% (exit rate) are leaving the website. On each of the three web pages (homepage, search, and program page) *similar search keywords* are used to search for programs, as illustrated in Figure 12.14.

Download Resource - Pocket Code Application Users

Figure 12.15 illustrates the user journey of the audience with the intent to download a media resource via the Pocket Code application. The user journey uses the *looks library* web page as the landing page, which is the second most frequented landing page of the Pocket Code application's users as stated in Table 12.5. Correspondingly, four of the five most frequented web pages are web pages of the Pocket Code library.

64% of the users of the looks library landing page are *downloading a resource* for the program creation. Additionally, 18.5% (bounce rate) of the visitors are leaving the website *before downloading a resource*, and 51% (exit rate) are terminating the session *after downloading* a resource.

Visitors are navigating from the looks web page to the sounds library 71.1%, the homepage 8.85%, the program page 52.8%, and the search web page 6.96%. Used search keywords on the looks web page are *Ball, Button, Cookie, Shop, Start, Car, Play, Ufo, Wasser, Background*. In opposition, search keywords on the sounds web page are names of sounds, such as *Horror, Bird, Click, or Dog*. Consequently, internal search keywords used on the Pocket Code library web pages are corresponding to the type of the displayed resource.

In the common user journey, the visitor navigates from the homepage to a program page, and revisits the homepage, as described in section 12.4.4. In contrast, users of the Pocket Code application are navigating from resource page to resource page. However, the usage of the website content of the same web pages, such as the homepage, search page, or program page, is similar (comparison of Figure 12.14 and Figure 12.15). 88.03% of the *sounds library*

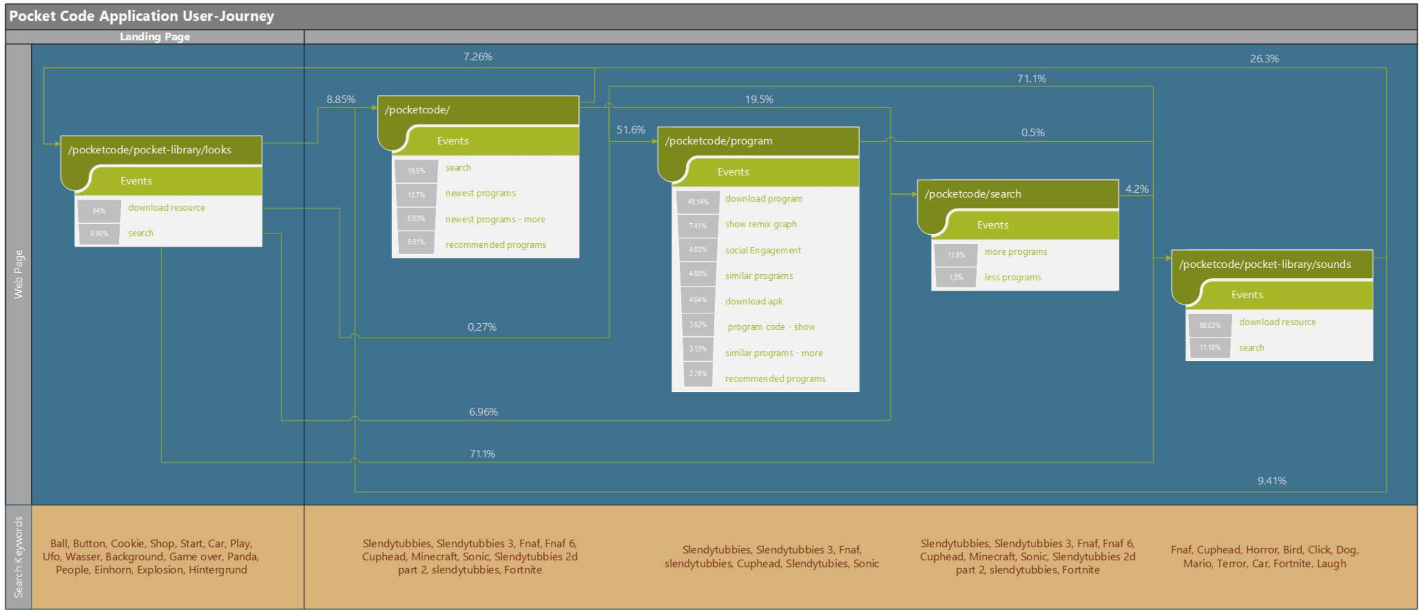

Pocket Code Application User-Journey
Landing Page
Web Page
Search Keywords
/pocketcode/pocket-library/looks
Events
64% download resource
6.96% search
8.85%
7.26%
/pocketcode/
Events
19.5% search
13.7% newest programs
6.93% newest programs - more
6.91% recommended programs
51.6%
19.5%
/pocketcode/program
Events
48.14% download program
7.41% show remix graph
4.63% social Engagement
4.60% similar programs
4.04% download apk
3.62% program code - show
3.13% similar programs - more
2.78% recommended programs
71.1%
0.5%
/pocketcode/search
Events
11.8% more programs
1.3% less programs
4.2%
26.3%
/pocketcode/pocket-library/sounds
Events
88.02% download resource
11.19% search
0,27%
6.96%
71.1%
9.41%
Ball, Button, Cookie, Shop, Start, Car, Play, Ufo, Wasser, Background, Game over, Panda, People, Einhorn, Explosion, Hintergrund
Slendytubbies, Slendytubbies 3, Fnaf, Fnaf 6, Cuphead, Minecraft, Sonic, Slendytubbies 2d part 2, slendytubbies, Fortnite
Slendytubbies, Slendytubbies 3, Fnaf, slendytubbies, Cuphead, Slendytubbies, Sonic
Slendytubbies, Slendytubbies 3, Fnaf, Fnaf 6, Cuphead, Minecraft, Sonic, Slendytubbies 2d part 2, slendytubbies, Fortnite
Fnaf, Cuphead, Horror, Bird, Click, Dog, Mario, Terror, Car, Fortnite, Laugh

Figure 12.15.: User Journey - Catrobat Application - Catrobat Community Website

visitors are *downloading a resource*. If they cannot find a suitable resource, 9.30% of the visitors are using the search function to search for sounds. The most common search keywords on the sounds web page are Cuphead, Horror, Bird, Click, Dog, Mario, Terror, Car, Fortnite, or Laugh. 46.1% of the visitors are leaving the website *after downloading a resource*, and 22.7% are terminating the session *before downloading a resource*. Furthermore, 53.9% of the users, who have downloaded a resource, are visiting web pages, such as the *looks library* to 26.3% and the *homepage* to 9.41%.

12.4.5. Video Content

Event tracking supports the analysis of the viewed video content, as described in section 7.2. The following section gives an overview of the consumed video content of the Catrobat sharing website. The total amount of visitors that are consuming the video content of the Catrobat community website are:

- 2.62% of the entire audience
- 0.02% of the Pocket Code application
- 15.29% of the web browser users

As outlined only a small portion of the audience is consuming the video content of the website. Of the video content consumers, most visitors (15.29%) are using the web browser to watch the offered content. In contrast, users of the Pocket Code application are rarely using the video content (0.02%). Table 12.6 list the 10 most viewed videos. As in column *% of total sessions* stated, the total amount of visitors who are viewing a video is very low. The most frequently viewed video is watched in 0.75% of the sessions.

12.5. Postvisit Stage

The postvisit stage tries to capture the user satisfaction of the visitor after leaving the website. However, web analytics does not provide capturing tools to acquire information about the postvisit stage, because only interactions with the website are captured. Possible information sources for the postvisit

Table 12.6.: Top 10 most viewed Videos globally

	Video Name	Number of Views (% of total Events)	% of total Sessions
1	Pocket Code - create your own games, directly on your phone!	500 (34.55%)	0.75%
2	Postcard - New Program	219 (15.13%)	0.33%
3	Postcard - New Object	145 (10.02%)	0.22%
4	(1) Animate your Characters	135 (9.33%)	0.20%
5	Postcard - Bricks	131 (9.05%)	0.20%
6	Inclination sensors and collision detection	65 (4.49%)	0.09%
7	Postcard - upload and share	58 (4.01%)	0.09%
8	Postcard - voice mail	54 (3.73%)	0.08%
9	(3a) Create a moving landscape and control the Hatters' movement	50 (3.46%)	0.07%
10	(3b) Let the Hatter shoot the evil cards	40 (2.76%)	0.06%

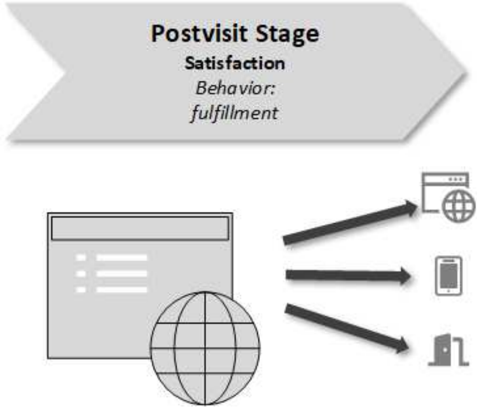


Figure 12.16.: Visit Stage. Based on [LV16; FKH15; SZ09]

stage are the Pocket Code application, news groups, email, social media interactions on the social media platform, the Google App store, partner websites, and YouTube.

13. Summary and Conclusion

In this final chapter, the preceding chapters are summarized, and the findings are outlined to form a conclusion about the individual topics discussed in this thesis. Concluding, this thesis presented a recommendation about further topics to be discussed.

13.1. Conclusion

At the initial stages of web analytics, the only available information source about the client-server communication was the server side gathered log file analysis. However, due to the limitation of the server-side gathered data, the log file analysis was succeeded by page tagging. Nowadays, the client-side script tagging is the most frequently used data capturing method of websites because of the detail of the gathered data, as outlined in chapter 4. Additionally, the JavaScript tracking code is under constant development and gets steadily improved to captures a holistic model of a website visit. Both data capturing methods have advantages and disadvantages in their capability to gather data about website visitors, which chapter 4 addresses. Event tracking supports the tracking of exclusive client-side interactions and user interactions on web pages via customized page tags and tracking code, as described in chapter 7.

Cookies are used to identify unique visitors and to track contiguous web page requests, as outlined in section 4.4. Web analytics tools are tracking these unique visitors by assigning a unique ID to each visitor. On each subsequent website visit, the cookie value is identified by the web analytics tool to recognize returning visitors. There are two types of cookies, first-party and third-party cookies. First-party cookies are usually used to save information for the web server about the currently viewed website, and third-party cookies are used to identify visitors across different domains. The behavior of web browser, plugins, and other network associated programs differs depending on the utilized cookie type, as discussed in section 4.4. For example, the deletion and rejection rate of third-party cookies is higher than the deletion and rejection rate of first-party cookies. Therefore, first-party cookies are mainly used to identify and track the clickstream of visitors on websites.

Two sources of data are available for the analysis of the website visitors' clickstream, the off-site metrics and the on-site metrics, as described in chapter 5. Off-site metrics collect the entire browsing history of Internet users. In contrast, on-site metrics are collecting data about the user interactions with the website, as introduced in section 5.2. Off-site metrics are providing information about the potential audience. In opposition, on-site metrics are analyzing the user behavior of the current audience. The on-site metrics are the primary information source for the analysis of website usage and are providing data about the visitor's clickstream, characteristics of the visitor, properties of the visit, and information about the viewed content.

The web reports are categorized into mainly four types, the traffic source metrics (accessibility), the visitor metrics (characterization), the behavior metrics (navigation), and the content metrics (design/content). Each metric group analyses a different aspect of a website visit, as outlined in section 5.3.1. Traffic source metrics investigate the arrival of the visitor, the visibility toward search engines, the interconnection with other websites, or the performance of advertisements and search keywords. The visitor metrics provide information about the visitor, the frequency and recency of the visits, the geographic location, or the device type. The metrics of the behavior category investigate properties of the visit, the web session duration, the page depth, the bounce rate, the navigation behavior, and the internal search function. The content metrics examine the viewed web pages of a session, the landing page, the exit page, the site speed, and the viewed multimedia content.

The trackable content of the website consists of page views, internal links, external links, and interactive content. By default the JavaScript page tag gathers only information about the internal navigation tracked by page views. This supports only a partial analysis of the behavior flow and clickstream depending on the interacted content of the website. The interactive content consists of partial updates of the web page content, links that are interacting with second- or third-party servers, or reactions to occurring events, such as the scrollbar tracking, or the tracking of the clicks. The tracking of external links and interactive content needs the implementation of customized tracking code, as described in chapter 7. Furthermore, page view tracking collects data about the navigation but not the interacted element. Customized tracking code needs to be added to identify the interacted element, as discussed in chapter 10.

Chapter 12 discusses the implementation of web analytics on the Pocket Code sharing website and evaluates the performance of the website based on selected reports and metrics. Section 12.4.4 displays the navigation behavior of two different user journeys, the journey of the entire audience and the journey of Pocket Code application users. However, the customer journey does not take technical aspects of the website in consideration, such as the page load time, and focuses entirely on the page path through the website. Hence, chapter 12 complement the customer journey with metrics that are investigating technical details, the characterization of visitors, and the used content of a website visit, as introduced in section 5.3.1.

The previsit stage analyses the origin of the website visitors through the investigation of the traffic sources. Based on the geographic location metrics, most of the website visitors are located in Austria, as outlined in section 12.2.1. Furthermore, the *direct traffic* is the main traffic source for the website with approximately 82.5% traffic share and 54,807 sessions (82.49%) in the observed time period. The direct traffic can be segmented in traffic from the Pocket Code application with 77.46% and traffic from other web browsers with 22.54%. Additional traffic sources are 14.27% social traffic, 3.07% referral traffic, and 0.17% traffic from organic search engines. This information suggests that search engines do not index the website correctly, and the main traffic source for the website is the Pocket Code application. Most of the social traffic 95.49% originates from YouTube. For a better social media integration of the audience, the engagement on other social media networks should be increased.

Web analytics reports can be used to determine information about the visiting behavior through the utilization of the *Count of Sessions* and *Days Since the Last Session* reports, as outlined in section 12.4.2. The *Count of Sessions* report states that a high percentage of the website visitors 51.15% visit the website only once. 20.99% of the audience visit the website two to three times, and only 27.86% of the returning visitors frequent it more than four times. Furthermore, the *Days Since the Last Session* reports yields that 81.18% of the returning visitors visit the website on the same day multiple times and then abandon the interaction with the website. Only 18.82% of the returning visitors (14.66% regular visitors and occasional visitors 4.16%) frequent the website after one day. Reasons for such a high abandonment rate can be that users do not find the expected content on the website or the web browsers are deleting the cookies and returning visitors cannot be identified.

Additional, web analytics supports the reconstruction of the user journey by utilizing the behavior flow graph and event tracking, as illustrated in section 12.4.4. The landing page metric of web analytics tools identifies the entry point to the website, as displayed in section 12.4.3. Figure 12.14 and Figure 12.15 in section 12.4.4 illustrates two frequented user journeys. These user journeys consist of the combination of the behavior flow graph (discussed in section 6.9), the event tracking (discussed in chapter 7), and the capturing of internal search queries (discussed in section 6.10).

This user journey comparison illustrates that the navigation path through the website depends on the landing page and user intent. Visitors from browsers are mostly landing on a *program* or the *help* page, and visitors from the Pocket Code application on the *homepage* or the *media library pages*, as stated in section 12.4.3. The navigation path through the website and search behavior is different between the two user journeys. The most frequented visitor path is the navigation from the homepage to a program page or search page, as illustrated in Figure 12.14. 48% of the program page visitors are downloading a Pocket Code program.

Website visitors of the Pocket Code media library web pages are mainly searching for resources to download. The bounce rate of 18.5% of the media library looks landing page is low, and only a small amount of visitors leaves the website without a single interaction. Most of the visitors are downloading a resource (64% of the visitors) and terminating the session afterward (51% of the visitors). Visitors that are further browsing the website are mainly visiting another web page of the Pocket Code library. 71.1% of the sounds library's visitors navigate to the looks library.

Additionally, the search keywords differ based on the content of the web page. Search keywords used on Pocket Code library web pages are corresponding to the content of the displayed resource. For example, the search keywords on the looks web page are Ball, Button, or Cookie, and on the sounds library are Horror, Bird, or Click. Search keywords used on the homepage, search page, and program page are corresponding to names of programs, for example, Slendytubbies, Slendytubbies 3, Fnaf, or Cuphead.

The video content of the website gets only used by 2.62% of the visitors. Segmented by the device, 15.29% visitors of web browsers are watching a video and only 0.02% of the Pocket Code application users. The mainly watched video is *"Pocket Code - create your own games, directly on your phone!"*. Reasons for the low viewing rate can be that the video content is not found, the video content is not engaging, or the visitor's expectations are not met.

13.2. Recommendation

In this section, a recommendation is given based on the results found in the clickstream analysis. As stated in the analysis of the traffic sources, the share of the visitors from organic search engines low. Since the traffic from search engines is low, an SEO should be conducted for the Pocket Code community website. An SEO optimizes the website for search engines, which subsequently increases the position of the Pocket Code community website in the search engine results to attract more visitors on this traffic channel. The page load time report resulted that the average page load time for the website is 7.15 seconds. As in section 6.14 stated, the page load time should not exceed two to three seconds. As a consequence, the code of the website should be optimized to reduce the page load time.

The video content of the website is rarely used, as stated in section 12.4.5. Only 2.62% of the entire audience views a offered video and the most seen video is the promotion video, *"Pocket Code - create your own games, directly on* *your phone!"*. Traffic from YouTube is a main traffic source, stated in section 12.3.1. Consequently, website visitors are highly video affine. However, the number of video views is low. This suggests that either the website videos are not found, or the content of the videos is not appealing enough. The optimization of the visitor path to the videos and an adjustment of the video content for the visitor needs should increase their performance.

The tracking of the internal search keyword resulted that visitors are also searching for sounds and looks. These items identified by the search keyword tracking can be added to the media library to offer developers resources of their interest. Web analytics is an iterative process of measurement and optimization. As a result, attributes of the website need to be quantified to measure their changes in each iteration. Such quantified outcomes are KPIs, PIs, Conversion Rates, CTRs, et. cetera and need to be defined depending on the business goals of the website. Only by comparing the former state and the current state of the website, performance changes are identifiable.

13.3. Further Work

For a detailed analysis of the website's performance, it is necessary to define outcomes, as described in section *Web Analytics* 3.3, such as Key Performance Indicators, Conversion Rates, CTRs and so forth. A topic of interest is the identification and definition of these outcomes for the Pocket Code community website. Furthermore, the monitoring of these outcomes support the identification of visitor behavior changes, as well as, the measurement of website design optimizations, the performance of newly implemented features, the influence of the Search Engine Optimization (SEO), or offline and online marketing campaign. The segmentation of the audience supports the identification of differences in their user behavior. Topics of interest are the algorithms for this segmentation (k-means) and subsequently the identification of their differences.

Appendix

Appendix A.

Tables

A.1. Reports

Table A.1.: Traffic Sources - Catrobat Community Website

	Acquisition			Behavior		
Source	Users	New Users	Sessions	Bounce Rate	Pages/ Session	Avg. Session Duration
Total	36,938	33,985	66,438	28.38%	5.31	00:06:57
(direct)	29,882	28,063	54,807	24.29%	5.86	00:07:46
youtube.com	5,188	4,469	8,707	48.93%	2.55	00:03:05
f2games.jimdo.com	320	133	485	54.64%	2.67	00:02:09
catrobat.org	274	211	476	32.35%	4.71	00:05:38
m.youtube.com	274	239	344	47.38%	2.30	00:02:30
en.scratch-wiki.info	174	172	196	42.35%	2.97	00:02:30
away.vk.com	156	122	296	44.93%	3.25	00:03:26
code.org	138	137	157	42.68%	2.54	00:01:31
scolartic.com	76	77	80	70.00%	1.76	00:00:53
gamejolt.com	57	52	64	35.94%	2.08	00:03:44
scratch-dach.info	30	23	58	32.76%	5.03	00:04:54
google	27	12	110	39.09%	3.74	00:03:55
m.facebook.com	24	24	36	38.89%	2.58	00:03:10
best-deal-hdd.pro	20	20	20	100.00%	1.00	00:00:00
github.com	19	13	43	27.91%	2.44	00:01:57
classroom.google.com	16	10	98	44.90%	1.05	00:01:09
l.facebook.com	15	14	16	81.25%	1.25	00:00:22
hourofcode.com	13	13	18	22.22%	2.44	00:03:21
scratch-ru.info	12	10	18	61.11%	2.56	00:02:28
alicegamejam.com	11	8	16	56.25%	3.06	00:01:00

Table A.1 continued from previous page

developer.catrobat.org	9	8	9	44.44%	4.11	00:03:11
galaxygamejam.com	9	7	24	25.00%	3.67	00:03:29
eartined.coursevo.com	8	8	21	85.71%	1.14	00:00:21
programamos.es	8	7	12	41.67%	2.42	00:02:01
restorecosm.bid	8	7	14	35.71%	4.64	00:06:33
sites.google.com	8	5	16	12.50%	3.12	00:08:43
codein.withgoogle.com	7	5	9	44.44%	1.44	00:00:35
forum.makeblock.com	7	7	7	28.57%	3.29	00:01:58
kidscodecs.com	6	6	7	14.29%	2.71	00:01:18
l.instagram.com	6	6	8	37.50%	2.38	00:00:51
mail.google.com	6	1	23	34.78%	3.39	00:05:30
plus.url.google.com	6	1	33	27.27%	4.85	00:02:40
en.wikipedia.org	5	5	5	20.00%	2.40	00:02:07
medien-in-die-schule.de	5	5	5	40.00%	2.60	00:00:39
recitmst.qc.ca	5	5	5	60.00%	1.60	00:00:27
blog.ozobot.com	4	3	11	63.64%	2.36	00:00:28
catrobatblog .wordpress.com	4	2	7	42.86%	1.86	00:01:16
facebook.com	4	4	4	25.00%	1.25	00:00:06
gamewizards.nl	4	4	4	0.00%	2.75	00:00:34
t.co	4	2	4	0.00%	3.25	00:03:47
womo.ua	4	4	8	62.50%	1.50	00:01:18
allyouneediscode.eu	3	3	6	33.33%	9.00	00:07:41
baidu	3	3	4	0.00%	3.25	00:01:37

Table A.1 continued from previous page

codigo21.educacion .navarra.es	3	2	3	33.33%	2.67	00:02:23
de.wikipedia.org	3	2	3	66.67%	1.33	00:00:01
moon.isvery.cool	3	3	3	66.67%	1.00	00:00:02
ru.scratch.wikia.com	3	3	4	50.00%	3.00	00:03:24
bloglenovo.es	2	2	2	0.00%	2.50	00:01:00
cafe.naver.com	2	2	3	33.33%	3.00	00:01:26
coderdojospqr.it	2	2	2	100.00%	1.00	00:00:00
codeweek.eu	2	1	3	66.67%	1.33	00:00:04
coursera.org	2	2	2	50.00%	1.00	00:00:19
en.m.wikipedia.org	2	2	2	100.00%	1.00	00:00:00
l.messenger.com	2	1	2	50.00%	5.50	00:03:17
newtonew.com	2	2	2	0.00%	5.00	00:05:58
oshl.edu.umh.es	2	2	2	0.00%	2.50	00:16:05
padlet.com	2	2	2	50.00%	3.00	00:00:11
researchgate.net	2	2	2	50.00%	2.00	00:01:51
sbox.edu.uni-graz.at	2	1	2	50.00%	1.50	00:01:23
4primariamatematicas .blogspot.com.es	1	1	1	0.00%	8.00	00:00:52
app.schoology.com	1	1	1	100.00%	1.00	00:00:00
ashgary.itch.io	1	1	1	0.00%	1.00	00:00:03
catrobat-scratch2 .ist.tugraz.at	1	0	1	100.00%	1.00	00:00:00
catrobat.slack.com	1	0	3	66.67%	1.00	00:00:04
courses.mc3.edu	1	1	1	100.00%	1.00	00:00:00

Table A.1 continued from previous page

deref-gmx.net	1	0	1	0.00%	5.00	00:02:19
elearning.lsr-noe.gv.at	1	0	1	0.00%	8.00	00:06:06
fxp.co.il	1	1	1	100.00%	1.00	00:00:00
gertjanschaap-wordpress-com.cdn.ampproject.org	1	0	1	100.00%	1.00	00:00:00
hdvidzpros.com	1	0	1	0.00%	3.00	00:00:09
hmtcre.simdif.com	1	0	13	38.46%	6.54	00:08:24
jugendprogrammiert .weebly.com	1	1	1	100.00%	1.00	00:00:00
livebinders.com	1	1	1	0.00%	3.00	00:00:07
lm.facebook.com	1	0	6	66.67%	2.67	00:00:15
localhost:44117	1	1	1	0.00%	3.00	00:21:07
localhost:63342	1	1	1	100.00%	1.00	00:00:00
loganemsinger.wixsite.com	1	0	1	0.00%	9.00	00:06:36
m.poczta.onet.pl	1	0	3	100.00%	1.00	00:00:00
microsoftfanon.wikia.com	1	1	1	100.00%	1.00	00:00:00
no1leftbehind.eu	1	0	2	50.00%	2.50	00:04:20
ok.ru	1	1	1	0.00%	3.00	00:00:29
oswajamyprogramowanie .edu.pl	1	1	1	0.00%	1.00	00:00:27
outlook.live.com	1	0	2	0.00%	8.00	00:06:46
pinterest.com	1	1	3	66.67%	1.33	00:05:49
pinterest.de	1	1	1	0.00%	2.00	00:00:11
poczta.wp.pl	1	0	1	0.00%	7.00	00:01:22
quirktools.com	1	0	13	30.77%	3.46	00:04:18

Table A.1 continued from previous page

robertpainsi.localhost.io	1	0	1	0.00%	1.00	00:00:05
sciencefocus.com	1	1	1	0.00%	2.00	00:00:18
sharelatex.com	1	1	1	0.00%	6.00	00:00:45
slideshare.net	1	1	1	0.00%	5.00	00:00:41
storage.googleapis.com	1	0	14	35.71%	4.50	00:10:01
teendeveloper.simdif.com	1	1	2	50.00%	1.50	00:04:55
tibs.at	1	0	1	0.00%	4.00	00:00:18
unascuola.it	1	1	1	0.00%	2.00	00:00:36
webmail.tugraz.at	1	0	14	14.29%	6.50	00:05:42
whirlpool.com.ar	1	1	1	100.00%	1.00	00:00:00
wiki.scratch.mit.edu	1	0	2	50.00%	2.50	00:06:31
yandex	1	1	1	0.00%	2.00	00:00:31
yandex.ru	1	1	1	0.00%	6.00	00:09:06
yandex.ua	1	1	1	0.00%	18.00	00:05:29
ziggo.nl	1	1	1	100.00%	1.00	00:00:00

Table A.2.: Referral Traffic - Catrobat Community Website

	Acquisition			Behavior		
Source	Users	New Users	Sessions	Bounce Rate	Pages/ Session	Avg. Session Duration
Total	1,330	1,010	2,038	43.62%	3.18	00:03:12
f2games.jimdo.com	320	133	485	54.64%	2.67	00:02:09
catrobat.org	274	211	476	32.35%	4.71	00:05:38
en.scratch-wiki.info	174	172	196	42.35%	2.97	00:02:30
code.org	138	137	157	42.68%	2.54	00:01:31
scolartic.com	76	77	80	70.00%	1.76	00:00:53
gamejolt.com	57	52	64	35.94%	2.08	00:03:44
scratch-dach.info	30	23	58	32.76%	5.03	00:04:54
best-deal-hdd.pro	20	20	20	100.00%	1.00	00:00:00
github.com	19	13	43	27.91%	2.44	00:01:57
classroom.google.com	16	10	98	44.90%	1.05	00:01:09
hourofcode.com	13	13	18	22.22%	2.44	00:03:21
scratch-ru.info	12	10	18	61.11%	2.56	00:02:28
alicegamejam.com	11	8	16	56.25%	3.06	00:01:00
developer.catrobat.org	9	8	9	44.44%	4.11	00:03:11
galaxygamejam.com	9	7	24	25.00%	3.67	00:03:29
eartined.coursevo.com	8	8	21	85.71%	1.14	00:00:21
programamos.es	8	7	12	41.67%	2.42	00:02:01
restorecosm.bid	8	7	14	35.71%	4.64	00:06:33
sites.google.com	8	5	16	12.50%	3.12	00:08:43
codein.withgoogle.com	7	5	9	44.44%	1.44	00:00:35

Table A.2 continued from previous page

forum.makeblock.com	7	7	7	28.57%	3.29	00:01:58
kidscodecs.com	6	6	7	14.29%	2.71	00:01:18
mail.google.com	6	1	23	34.78%	3.39	00:05:30
en.wikipedia.org	5	5	5	20.00%	2.40	00:02:07
medien-in-die-schule.de	5	5	5	40.00%	2.60	00:00:39
recitmst.qc.ca	5	5	5	60.00%	1.60	00:00:27
blog.ozobot.com	4	3	11	63.64%	2.36	00:00:28
gamewizards.nl	4	4	4	0.00%	2.75	00:00:34
womo.ua	4	4	8	62.50%	1.50	00:01:18
allyouneediscode.eu	3	3	6	33.33%	9.00	00:07:41
codigo21.educacion .navarra.es	3	2	3	33.33%	2.67	00:02:23
de.wikipedia.org	3	2	3	66.67%	1.33	00:00:01
moon.isvery.cool	3	3	3	66.67%	1.00	00:00:02
bloglenovo.es	2	2	2	0.00%	2.50	00:01:00
coderdojospqr.it	2	2	2	100.00%	1.00	00:00:00
codeweek.eu	2	1	3	66.67%	1.33	00:00:04
coursera.org	2	2	2	50.00%	1.00	00:00:19
en.m.wikipedia.org	2	2	2	100.00%	1.00	00:00:00
l.messenger.com	2	1	2	50.00%	5.50	00:03:17
newtonew.com	2	2	2	0.00%	5.00	00:05:58
oshl.edu.umh.es	2	2	2	0.00%	2.50	00:16:05
padlet.com	2	2	2	50.00%	3.00	00:00:11
sbox.edu.uni-graz.at	2	1	2	50.00%	1.50	00:01:23
app.schoology.com	1	1	1	100.00%	1.00	00:00:00

Table A.2 continued from previous page

ashgary.itch.io	1	1	1	0.00%	1.00	00:00:03
catrobat-scratch2 .ist.tugraz.at	1	0	1	100.00%	1.00	00:00:00
catrobat.slack.com	1	0	3	66.67%	1.00	00:00:04
courses.mc3.edu	1	1	1	100.00%	1.00	00:00:00
deref-gmx.net	1	0	1	0.00%	5.00	00:02:19
elearning.lsr-noe.gv.at	1	0	1	0.00%	8.00	00:06:06
fxp.co.il	1	1	1	100.00%	1.00	00:00:00
gertjanschaap-wordpress -com.cdn.ampproject.org	1	0	1	100.00%	1.00	00:00:00
hdvidzpros.com	1	0	1	0.00%	3.00	00:00:09
hmtcre.simdif.com	1	0	13	38.46%	6.54	00:08:24
livebinders.com	1	1	1	0.00%	3.00	00:00:07
localhost:44117	1	1	1	0.00%	3.00	00:21:07
localhost:63342	1	1	1	100.00%	1.00	00:00:00
loganemsinger.wixsite.com	1	0	1	0.00%	9.00	00:06:36
m.poczta.onet.pl	1	0	3	100.00%	1.00	00:00:00
no1leftbehind.eu	1	0	2	50.00%	2.50	00:04:20
ok.ru	1	1	1	0.00%	3.00	00:00:29
oswajamyprogramowanie .edu.pl	1	1	1	0.00%	1.00	00:00:27
outlook.live.com	1	0	2	0.00%	8.00	00:06:46
poczta.wp.pl	1	0	1	0.00%	7.00	00:01:22
quirktools.com	1	0	13	30.77%	3.46	00:04:18
robertpainsi.localhost.io	1	0	1	0.00%	1.00	00:00:05

Table A.2 continued from previous page

sciencefocus.com	1	1	1	0.00%	2.00	00:00:18
sharelatex.com	1	1	1	0.00%	6.00	00:00:45
storage.googleapis.com	1	0	14	35.71%	4.50	00:10:01
teendeveloper.simdif.com	1	1	2	50.00%	1.50	00:04:55
tibs.at	1	0	1	0.00%	4.00	00:00:18
unascuola.it	1	1	1	0.00%	2.00	00:00:36
webmail.tugraz.at	1	0	14	14.29%	6.50	00:05:42
whirlpool.com.ar	1	1	1	100.00%	1.00	00:00:00
wiki.scratch.mit.edu	1	0	2	50.00%	2.50	00:06:31
yandex.ru	1	1	1	0.00%	6.00	00:09:06
yandex.ua	1	1	1	0.00%	18.00	00:05:29
ziggo.nl	1	1	1	100.00%	1.00	00:00:00

Table A.3.: Browser - Catrobat Community Website

	Acquisition			Behavior		
Browser	Users	New Users	Sessions	Bounce Rate	Pages/ Session	Avg. Session Duration
Total	36,717	33,985	66,438	28.38%	5.31	00:06:57
Catrobat	20,774	19,213	42,454	17.20%	6.78	00:09:26
Chrome	8,479	7,540	14,685	50.59%	2.76	00:03:06
Safari	4,812	4,797	4,895	37.36%	2.35	00:00:13
Samsung Internet	786	698	1,398	55.01%	2.69	00:03:06
Android Webview	698	652	1,138	57.38%	2.28	00:02:33
Firefox	269	245	523	34.61%	4.66	00:05:08
Android Browser	258	231	399	62.66%	2.25	00:02:38
UC Browser	115	111	166	57.23%	2.43	00:02:19
YaBrowser	112	105	146	39.04%	2.98	00:02:56
Edge	96	91	138	47.83%	3.35	00:04:31
Opera	78	71	135	47.41%	2.79	00:04:34
Internet Explorer	75	77	98	37.76%	4.76	00:05:30
Amazon Silk	74	70	146	47.26%	2.84	00:04:16
Puffin	46	41	60	41.67%	3.53	00:02:17
Opera Mini	25	24	32	56.25%	3.66	00:04:21
Safari (in-app)	12	11	12	50.00%	3.33	00:01:32
Mozilla Compatible Agent	5	5	8	50.00%	5.50	00:03:37
Coc Coc	1	1	3	0.00%	7.00	00:03:56
Maxthon	1	1	1	100.00%	1.00	00:00:00
Playstation Vita Browser	1	1	1	100.00%	1.00	00:00:00

Table A.4.: Days Since Last Session - All Traffic - Catrobat Community Website

Days Since Last Session	Sessions	Page Views
0	53,939	282,491
1	3,132	18,815
2	1,401	7,832
3	947	5,400
4	752	4,373
5	649	3,849
6	812	5,475
7	462	2,452
8-14	1,591	8,647
15-30	1,412	6,996
31-60	749	3,639
61-120	418	1,716
121-364	174	828
	66,438	352,513

Table A.6.: Count of Sessions - Catrobat Community Website

Count of Sessions	Segment	Sessions
2	Browser - Catrobat	824
	Direct Traffic	1008
	Referral Traffic	178
	Social Media Traffic	136
3	Browser - Catrobat	421
	Direct Traffic	514
	Referral Traffic	78
	Social Media Traffic	59
4	Browser - Catrobat	309
	Direct Traffic	363
	Referral Traffic	37
	Social Media Traffic	27
5	Browser - Catrobat	220

Table A.6 continued from previous page

	Direct Traffic	248
	Referral Traffic	25
	Social Media Traffic	17
6	Browser - Catrobat	159
	Direct Traffic	174
	Referral Traffic	17
	Social Media Traffic	12
7	Browser - Catrobat	129
	Direct Traffic	138
	Referral Traffic	19
	Social Media Traffic	16
8	Browser - Catrobat	98
	Direct Traffic	107
	Referral Traffic	15
	Social Media Traffic	12
9-14	Browser - Catrobat	328
	Direct Traffic	359
	Referral Traffic	28
	Social Media Traffic	10
15-25	Browser - Catrobat	342
	Direct Traffic	359
	Referral Traffic	28
	Social Media Traffic	10
26-50	Browser - Catrobat	332
	Direct Traffic	336
	Referral Traffic	12
	Social Media Traffic	3
51-100	Browser - Catrobat	94
	Direct Traffic	94
	Referral Traffic	8
	Social Media Traffic	5
101-200	Browser - Catrobat	15
	Direct Traffic	17
	Referral Traffic	8

Table A.6 continued from previous page

	Social Media Traffic	8
201+	Browser - Catrobat	18
	Direct Traffic	20
	Referral Traffic	12
	Social Media Traffic	5

Table A.7.: Days Since Last Session - Catrobat Community Website

Days Since Last Sessions	Segment	Sessions
1	Browser - Catrobat	321
	Direct Traffic	368
	Referral Traffic	43
	Social Media Traffic	37
2	Browser - Catrobat	157
	Direct Traffic	183
	Referral Traffic	31
	Social Media Traffic	21
3	Browser - Catrobat	105
	Direct Traffic	120
	Referral Traffic	20
	Social Media Traffic	16
4	Browser - Catrobat	65
	Direct Traffic	78
	Referral Traffic	10
	Social Media Traffic	9
5	Browser - Catrobat	56
	Direct Traffic	63
	Referral Traffic	8
	Social Media Traffic	6
6	Browser - Catrobat	61
	Direct Traffic	63
	Referral Traffic	11

Table A.7 continued from previous page

	Social Media Traffic	10
7	Browser - Catrobat	48
	Direct Traffic	56
	Referral Traffic	6
	Social Media Traffic	2
8-14	Browser - Catrobat	110
	Direct Traffic	141
	Referral Traffic	24
	Social Media Traffic	18
15-30	Browser - Catrobat	94
	Direct Traffic	112
	Referral Traffic	24
	Social Media Traffic	16
31-60	Browser - Catrobat	49
	Direct Traffic	66
	Referral Traffic	19
	Social Media Traffic	10
61-120	Browser - Catrobat	26
	Direct Traffic	38
	Referral Traffic	9
	Social Media Traffic	8
51-100	Browser - Catrobat	94
	Direct Traffic	94
	Referral Traffic	8
	Social Media Traffic	5
121+	Browser - Catrobat	11
	Direct Traffic	14
	Referral Traffic	5
	Social Media Traffic	5

Table A.5.: Count of Sessions - All Traffic - Catrobat Community Website

Count of Sessions	Sessions	Page Views
1	33,985	150,173
2	9,173	47,310
3	4,773	28,988
4	3,025	19,234
5	2,145	13,580
6	1,635	10,283
7	1,280	8,437
8	1,035	6,954
9-14	3,583	25,329
15-25	2,548	17,892
26-50	1,769	12,902
51-100	946	6,951
101-200	364	3,404
201+	177	1,076
	66,438	352,513

Table A.8.: Landing Pages - Catrobat Community Website

	Landing Page	Segment	Sessions	% New Sessions	% New Users	Bounce Rate	Pages/ Session	Avg. Session Duration
1	/pocketcode/	All Users	24,116	41.22%	9,940	14.97%	8.16	00:08:19
		Catrobat App	22,341	39.96%	8,927	13.46%	8.42	00:08:35
		Browser	1,775	57.07%	1,013	33.86%	4.77	00:04:59
2	/pocketcode/program/	All Users	12,163	53.34%	6,488	48.91%	2.61	00:03:07
		Catrobat App	1,118	62.88%	703	43.74%	4.03	00:04:23
		Browser	11,045	52.38%	5,785	49.43%	2.47	00:03:00
3	/pocketcode/pocket-library/looks	All Users	7,794	36.58%	2,851	18.53%	5.26	00:13:05
		Catrobat App	7,793	36.58%	2,851	18.52%	5.26	00:13:05
		Browser	1	0.00%	0	100.00%	1.00	00:00:00
4	/pocketcode/help/	All Users	5,173	59.71%	3,089	58.63%	2.85	00:02:50
		Catrobat App	60	63.33%	38	56.67%	6.22	00:11:06
		Browser	5,113	59.67%	3,051	58.65%	2.81	00:02:44
5	/pocketcode/pocket-library/sounds	All Users	4,143	57.62%	2,387	25.25%	4.35	00:07:31
		Catrobat App	4,143	57.62%	2,387	25.25%	4.35	00:07:31
		Browser	0	0.00%	0	0.00%	0.00	00:00:00
6	/pocketcode/pocket-library/backgrounds-portrait	All Users	3,628	58.93%	2,138	14.75%	5.05	00:11:26
		Catrobat App	3,627	58.95%	2,138	14.72%	5.05	00:11:27
		Browser	1	0.00%	0	100.00%	1.00	00:00:00
7	/pocketcode/pocket-library/backgrounds-landscape	All Users	2,687	70.12%	1,884	23.11%	4.67	00:08:47
		Catrobat App	2,687	70.12%	1,884	23.11%	4.67	00:08:47
		Browser	0	0.00%	0	0.00%	0.00	00:00:00
8	/pocketcode/search/	All Users	349	8.88%	31	22.35%	7.34	00:08:36
		Catrobat App	178	12.36%	22	2.25%	9.35	00:10:39
		Browser	171	5.26%	9	43.27%	5.24	00:06:28
9	/pocketcode/stepByStep	All Users	258	31.40%	81	67.44%	2.22	00:03:13
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	258	31.40%	81	67.44%	2.22	00:03:13
10	/pocketcode/resetting/request	All Users	222	68.92%	153	25.68%	7.69	00:08:03
		Catrobat App	218	69.72%	152	25.23%	7.78	00:08:12
		Browser	4	25.00%	1	50.00%	2.50	00:00:31
11	/pocketcode/login/	All Users	122	12.30%	15	50.82%	3.22	00:03:35
		Catrobat App	24	25.00%	6	4.17%	7.33	00:08:11
		Browser	98	9.18%	9	62.24%	2.21	00:02:27
12	/pocketcode/termsOfUse	All Users	91	54.95%	50	68.13%	2.37	00:02:33
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	91	54.95%	50	68.13%	2.37	00:02:33
13	/pocketcode/tutorialcards	All Users	73	23.29%	17	52.05%	3.45	00:03:54
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	73	23.29%	17	52.05%	3.45	00:03:54
14	/pocketcode/gaming-tutorials	All Users	70	14.29%	10	50.00%	2.87	00:04:37
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	70	14.29%	10	50.00%	2.87	00:04:37
15	/login	All Users	69	59.42%	41	28.99%	6.74	00:06:06
		Catrobat App	52	71.15%	37	30.77%	7.52	00:06:44
		Browser	17	23.53%	4	23.53%	4.35	00:04:11
16	/create@school/	All Users	62	20.97%	13	11.29%	7.08	00:05:20
		Catrobat App	62	20.97%	13	11.29%	7.08	00:05:20
		Browser	0	0.00%	0	0.00%	0.00	00:00:00
17	/pocketgalaxy/	All Users	59	93.22%	55	35.59%	2.32	00:01:46
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00

Table A.8 continued from previous page

		Browser	59	93.22%	55	35.59%	2.32	00:01:46
18	/pocketcode/starter-programs	All Users	54	33.33%	18	51.85%	3.09	00:01:41
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	54	33.33%	18	51.85%	3.09	00:01:41
19	/pocketcode/gaming-tutorials/1	All Users	51	5.88%	3	82.35%	1.59	00:03:52
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	51	5.88%	3	82.35%	1.59	00:03:52
20	/pocketcode/gaming-tutorials/2	All Users	46	2.17%	1	73.91%	1.87	00:04:00
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	46	2.17%	1	73.91%	1.87	00:04:00
21	/pocketcode/tag/search/	All Users	28	7.14%	2	64.29%	2.43	00:02:38
		Catrobat App	4	0.00%	0	25.00%	4.75	00:06:08
		Browser	24	8.33%	2	70.83%	2.04	00:02:03
22	/luna/	All Users	26	23.08%	6	42.31%	3.50	00:02:21
		Catrobat App	18	33.33%	6	33.33%	3.89	00:02:22
		Browser	8	0.00%	0	62.50%	2.62	00:02:19
23	/index.html	All Users	20	100.00%	20	100.00%	1.00	00:00:00
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	20	100.00%	20	100.00%	1.00	00:00:00
24	/pocketcode/gaming-tutorials/4	All Users	20	0.00%	0	75.00%	1.80	00:00:57
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	20	0.00%	0	75.00%	1.80	00:00:57
25	/pocketcode/profile/	All Users	20	10.00%	2	55.00%	3.50	00:01:39
		Catrobat App	5	40.00%	2	0.00%	7.20	00:01:47
		Browser	15	0.00%	0	73.33%	2.27	00:01:36
26	/pocketcode/program/44987	All Users	20	15.00%	3	85.00%	1.15	00:01:09
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	20	15.00%	3	85.00%	1.15	00:01:09
27	/pocketcode/pocket-library/backgrounds	All Users	13	69.23%	9	30.77%	3.15	00:07:28
		Catrobat App	13	69.23%	9	30.77%	3.15	00:07:28
		Browser	0	0.00%	0	0.00%	0.00	00:00:00
28	/pocketcode/tutorial	All Users	13	84.62%	11	23.08%	3.62	00:04:50
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	13	84.62%	11	23.08%	3.62	00:04:50
29	/pocketcode/tutorialcards/1	All Users	11	9.09%	1	72.73%	1.36	00:01:58
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	11	9.09%	1	72.73%	1.36	00:01:58
30	/pocketcode/tutorialcards/11	All Users	11	0.00%	0	90.91%	1.09	00:00:56
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	11	0.00%	0	90.91%	1.09	00:00:56
31	/pocketcode/tutorialcards/6	All Users	11	0.00%	0	90.91%	1.27	00:00:47
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	11	0.00%	0	90.91%	1.27	00:00:47
32	/pocketcode/tutorialcards/8	All Users	11	18.18%	2	63.64%	1.73	00:02:11
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	11	18.18%	2	63.64%	1.73	00:02:11
33	/pocketcode/gaming-tutorials/3	All Users	10	0.00%	0	50.00%	5.00	00:02:35
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	10	0.00%	0	50.00%	5.00	00:02:35
34	/pocketcode/gaming-tutorials/5	All Users	10	0.00%	0	70.00%	2.00	00:02:03
		Catrobat App	0	0.00%	0	0.00%	0.00	00:00:00
		Browser	10	0.00%	0	70.00%	2.00	00:02:03
35	/pocketcode/register/	All Users	9	11.11%	1	22.22%	4.67	00:03:30
		Catrobat App	2	0.00%	0	0.00%	4.00	00:10:40

Table A.8 continued from previous page

			Browser	7	14.29%	1	28.57%	4.86	00:01:27

Bibliography

[AML07] M Abraham, CAMERON Meierhoefer, and ANDREW Lipsman. "The impact of cookie deletion on the accuracy of site-server and ad-server metrics: An empirical comScore study." In: 14 (2007), p. 2009 (cit. on pp. 22, 32, 44).

[ANF12] Sareh Aghaei, Mohammad Ali Nematbakhsh, and Hadi Khosravi Farsani. "Evolution of the world wide web: From WEB 1.0 TO WEB 4.0." In: *International Journal of Web & Semantic Technology* 3.1 (2012), p. 1 (cit. on p. 6).

[Aka10] Inc. Akamai Technologies. *Akamai – New Study Reveals the Impact of Travel Site Performance on Consumers*. June 2010. URL: `https://www.akamai.com/us/en/about/news/press/2010-press/new-study-reveals-the-impact-of-travel-site-performance-on-consumers.jsp` (visited on 05/10/2019) (cit. on p. 57).

[Ald06] SE Aldrich. "The Other Search: Making the Most of Site Search to Optimize the Total Customer Experience." In: *Patricia Seybold Group* (2006) (cit. on p. 55).

[Ana19a] Google Analytics. *Analytics Help – About Demographics and Interests*. May 2019. URL: `https://support.google.com/analytics/answer/2799357?hl=en` (visited on 05/10/2019) (cit. on p. 96).

[Ana19b] Google Analytics. *Google Analytics – Benefits of User-ID*. Apr. 2019. URL: `https://support.google.com/analytics/answer/3123663` (visited on 05/10/2019) (cit. on pp. 44, 45).

[Ana18] Google Analytics. *Google Analytics – Google Analytics Cookie Usage on Websites*. Aug. 2018. URL: `https://developers.google.com/analytics/devguides/collection/analyticsjs/cookie-usage` (visited on 05/10/2019) (cit. on p. 43).

[Ana19c] Google Analytics. *Google Analytics – Tracking*. Mar. 2019. URL: `https://developers.google.com/analytics/devguides/collection/analyticsjs/` (visited on 05/10/2019) (cit. on pp. 4, 25, 27, 42, 43).

[And06] Chris Anderson. *The long tail: How endless choice is creating unlimitted demand*. 2006 (cit. on p. 61).

[Aye+11] Mika D Ayenson et al. "Flash cookies and privacy II: Now with HTML5 and ETag respawning." In: *SSRN Electronic Journal* (2011) (cit. on p. 45).

[BHF12] Christian Banse, Dominik Herrmann, and Hannes Federrath. "Tracking users on the internet with behavioral patterns: Evaluation of its practical feasibility." In: *IFIP International Information Security Conference*. Springer. 2012, pp. 235–248 (cit. on p. 36).

[Bea13] Michael Beasley. *Practical web analytics for user experience: How analytics can help you understand your users*. Newnes, 2013 (cit. on pp. 4, 14, 17, 19, 20, 22–26, 28, 29, 32, 36, 37, 41, 42, 44, 45, 50–56, 58, 59, 63, 64, 66).

[BG15] Ivan Bekavac and Daniela Garbin Praničević. "Web analytics tools and web metrics tools: An overview and comparative analysis." In: *Croatian Operational Research Review* 6.2 (2015), pp. 373–386 (cit. on pp. 38, 40, 46–48, 52).

[BS13] Bhavna Beri and Parminder Singh. "Web analytics: Increasing website's usability and conversion rate." In: *International Journal of Computer Applications* 72.6 (2013) (cit. on pp. 14, 39, 42, 45).

[BA17] Gaël Bernard and Periklis Andritsos. "A Process Mining Based Model for Customer Journey Mapping." In: (2017) (cit. on pp. 110–112, 114).

[Ber89] Timothy J Berners-Lee. *Information management: A proposal*. Tech. rep. 1989 (cit. on p. 6).

[BB98] Krishna Bharat and Andrei Broder. "A technique for measuring the relative size and overlap of public web search engines." In: *Computer Networks and ISDN systems* 30.1 (1998), pp. 379–388 (cit. on p. 36).

[BI14] Benjamin Billet and Valérie Issarny. "Dioptase: a distributed data streaming middleware for the future web of things." In: *Journal of Internet Services and Applications* 5.1 (2014), p. 13 (cit. on p. 6).

[BA03] Robert Blumberg and Shaku Atre. "The problem with unstructured data." In: *Dm Review* 13.42-49 (2003), p. 62 (cit. on p. 10).

[Bon12] John Bonini. *Impact – The 5 Most Important Website Statistics You Should Be Tracking*. Aug. 2012. URL: `https://web.archive.org/web/20151108190010/http://www.impactbnd.com/blog/the-5-most-important-website-statistics-you-should-be-tracking` (visited on 05/10/2019) (cit. on p. 42).

[BJ10] Danielle Booth and Bernard J Jansen. "A review of methodologies for analyzing websites." In: *Web technologies: Concepts, methodologies, tools, and applications*. IGI Global, 2010, pp. 145–166 (cit. on pp. 53–56).

[BP12] Sergey Brin and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." In: *Computer networks* 56.18 (2012), pp. 3825–3833 (cit. on p. 19).

[Bro02] Andrei Broder. "A taxonomy of web search." In: *ACM Sigir forum*. Vol. 36. 2. ACM. 2002, pp. 3–10 (cit. on p. 36).

[BHS11] Erik Brynjolfsson, Yu Hu, and Duncan Simester. "Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales." In: *Management Science* 57.8 (2011), pp. 1373–1386 (cit. on p. 61).

[BA07] Jason Burby and Shane Atchison. *Actionable web analytics: using data to make smart business decisions*. John Wiley & Sons, 2007 (cit. on pp. 3, 11, 12, 14, 16, 17, 32, 37).

[BBW+07] Jason Burby, Angie Brown, Web Analytics Association (WAA) Standards Committee, et al. "Web analytics definitions." In: *Washington DC: Web Analytics Association* (2007) (cit. on pp. 4, 11, 12, 22, 37, 39, 41, 42, 46, 50–52, 56, 59, 63).

[Cam14] Jennifer T. Campbell. *Web Design: Introductory*. Cengage Learning, 2014 (cit. on p. 80).

[Cas01] Manuel Castells. *The Internet Galaxy: Reflections on the Internet, Business, and Society*. Oxford University Press, Inc., 2001 (cit. on p. 6).

[Cat19] Catrobat. *Catrobat – Homepage*. May 2019. URL: https://www.catrobat.org/ (visited on 05/10/2019) (cit. on p. 73).

[CDN11] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. "An overview of business intelligence technology." In: *Communications of the ACM* 54.8 (2011), pp. 88–98 (cit. on pp. 2, 10, 11).

[CCS12] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. "Business intelligence and analytics: From big data to big impact." In: *MIS quarterly* 36.4 (2012), pp. 1165–1188 (cit. on pp. 2, 7–11).

[CB10] Jeungok Choi and Suzanne Bakken. "Web-based education for low-literate parents in Neonatal Intensive Care Unit: Development of a website and heuristic evaluation and usability testing." In: *International journal of medical informatics* 79.8 (2010), pp. 565–575 (cit. on p. 15).

[Cli12] Brian Clifton. *Advanced web metrics with Google Analytics*. John Wiley & Sons, 2012 (cit. on pp. 2, 4, 20, 22, 23, 25–28, 31, 32, 34, 35, 37, 43–46, 53–57, 66–68, 70, 71, 95).

[Cli15a] Brian Clifton. *The Insights Blog – 8 Recommendations For Choosing A Tool: on-site versus off-site analytics*. Aug. 2015. URL: https://brianclifton.com/blog/2015/08/20/on-site-versus-off-site-analytics-which-is-accurate/ (visited on 05/10/2019) (cit. on pp. 4, 34, 36, 37).

[Cli15b] Brian Clifton. *The Insights Blog – Improving The Web Using Data*. Aug. 2015. URL: https://brianclifton.com/blog/2015/08/20/improving-the-web-with-web-analytics/ (visited on 05/10/2019) (cit. on pp. 33, 34).

[Cli15c] Brian Clifton. *The Insights Blog – On-site Versus Off-site Web Analytics – 2. How They Work*. Aug. 2015. URL: https://brianclifton.com/blog/2015/08/20/on-site-versus-off-site-web-analytics/ (visited on 05/10/2019) (cit. on pp. 34–36).

[Coh13] Raphael Cohen-Almagor. "Internet history." In: *Moral, Ethical, and Social Dilemmas in the Age of Technology: Theories and Practice*. IGI Global, 2013, pp. 19–39 (cit. on p. 6).

[Cuk10] Kenneth Cukier. *Data, data everywhere: A special report on managing information*. Economist Newspaper, 2010 (cit. on pp. 8, 9).

[Cus01] Brian O Cusack. "Measurements on the Web: Where do they Lead." In: *Proceedings of the NACCQ*. 2001 (cit. on p. 30).

[DM13] Sergei N Dorogovtsev and José FF Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013 (cit. on p. 6).

[DW07] Abhijit Dubey and Dilip Wagle. "Delivering software as a service." In: *The McKinsey Quarterly* 6.2007 (2007), p. 2007 (cit. on p. 24).

[DR11] Esther Düweke and Stefan Rabsch. "Erfolgreiche Websites." In: *SEO, SEM, online-marketing, usability* 1 (2011) (cit. on pp. 62, 64).

[eBi17] eBizMba. *eBizMba – Top 15 Most Popular Search Engines — July 2017*. July 2017. URL: `http://www.ebizmba.com/articles/search-engines` (visited on 08/05/2017) (cit. on p. 36).

[End+08] Albrecht Enders et al. "The long tail of social networking.: Revenue models of social networking sites." In: *European Management Journal* 26.3 (2008), pp. 199–211 (cit. on p. 61).

[Eng07] Eric Enge. *Stone Temple Consulting Corporation – Web Analytics Shootout – Final Report – a 2011 Perspective*. Aug. 2007. URL: `https://www.stonetemple.com/web-analytics-shootout-final-report-a-2011-perspective/` (visited on 05/10/2019) (cit. on pp. 32, 43, 44).

[Eng16] Jonathan English. *Skeleton – 7 Key Video Metrics to Measure the Success of Your Content*. Nov. 2016. URL: `https://www.skeletonproductions.com/insights/video-metrics` (visited on 05/10/2019) (cit. on pp. 57, 58).

[FB13] Wei Fan and Albert Bifet. "Mining big data: current status, and forecast to the future." In: *ACM sIGKDD Explorations Newsletter* 14.2 (2013), pp. 1–5 (cit. on pp. 3, 8, 9).

[Fay+96] Usama M Fayyad et al. *Advances in knowledge discovery and data mining*. the MIT Press, 1996 (cit. on p. 11).

[Fie+99] Roy Fielding et al. *Hypertext transfer protocol–HTTP/1.1*. Tech. rep. 1999 (cit. on p. 31).

[FKH15] Asbjørn Følstad, Knut Kvale, and Ragnhild Halvorsrud. "Customer journey measures-State of the art research and best practices." In: *SINTEF Report A24488* (2015) (cit. on pp. 110, 111, 113, 118, 123, 133).

[Fom10] Max I Fomitchev. "How google analytics and conventional cookie tracking techniques overestimate unique visitors." In: *WWW* 10 (2010), pp. 1093–1094 (cit. on p. 20).

[GH15] Amir Gandomi and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." In: *International Journal of Information Management* 35.2 (2015), pp. 137–144 (cit. on pp. 3, 8–10).

[GR11] John Gantz and David Reinsel. "Extracting value from chaos." In: *IDC iview* 1142.2011 (2011), pp. 1–12 (cit. on p. 8).

[GH11] Carolin Gerlitz and Anne Helmond. "Hit, link, like and share. Organising the social and the fabric of the web." In: *Digital Methods Winter Conference Proceedings*. 2011, pp. 1–29 (cit. on pp. 70, 71).

[GH13] Carolin Gerlitz and Anne Helmond. "The like economy: Social buttons and the data-intensive web." In: *New Media & Society* 15.8 (2013), pp. 1348–1365 (cit. on p. 71).

[GM09] Manish Godse and Shrikant Mulik. "An approach for selecting software-as-a-service (SaaS) product." In: *Cloud Computing, 2009. CLOUD'09. IEEE International Conference on*. IEEE. 2009, pp. 155–158 (cit. on p. 24).

[GFH14] Yasemin Gokcen, Vahid Aghaei Foroushani, and A Nur Zincir Heywood. "Can we identify NAT behavior by analyzing Traffic Flows?" In: *2014 IEEE Security and Privacy Workshops*. IEEE. 2014, pp. 132–139 (cit. on p. 20).

[Goo19a] Google. *Google – About Events*. May 2019. URL: https://support.google.com/analytics/answer/1033068 (visited on 05/10/2019) (cit. on p. 70).

[Goo19b] Google. *Google – About Social plugins and interactions*. May 2019. URL: https://support.google.com/analytics/answer/6209874 (visited on 05/10/2019) (cit. on p. 71).

[Goo19c] Google. *Google Analytics – About Content Grouping*. May 2019. URL: https://support.google.com/analytics/answer/2853423?hl=en (visited on 05/10/2019) (cit. on p. 54).

[Goo19d] Google. *Google Analytics Developer – Event Tracking*. Mar. 2019. URL: https://developers.google.com/analytics/devguides/collection/analyticsjs/events (visited on 05/10/2019) (cit. on pp. 66, 70).

[Goo18] Google. *Google Analytics Developer – Page Tracking*. June 2018. URL: https://developers.google.com/analytics/devguides/collection/analyticsjs/pages (visited on 05/10/2019) (cit. on p. 66).

[HRC11] Richard Hanna, Andrew Rohm, and Victoria L Crittenden. "We're all connected: The power of the social media ecosystem." In: *Business horizons* 54.3 (2011), pp. 265–273 (cit. on p. 65).

[Har+13] Annemarie Harzl et al. "A Scratch-like visual programming system for Microsoft Windows Phone 8." In: *arXiv preprint arXiv:1310.1390* (2013) (cit. on p. 75).

[HMP09] Layla Hasan, Anne Morris, and Steve Probets. "Using Google Analytics to evaluate the usability of e-commerce sites." In: *Human centered design* (2009), pp. 697–706 (cit. on pp. 2, 11, 16, 23, 24, 26, 29, 37, 53).

[Has07] Marco Hassler. "Web Analytics - Zielorientierte Nutzung zur Erfolgssteigerung." In: *Whitepaper, namics, St.Gallen* (2007) (cit. on pp. 2, 4, 54–56, 58, 63, 71).

[Has12] Marco Hassler. *Web analytics: Metriken auswerten, Besucherverhalten verstehen, Website optimieren*. MITP-Verlags GmbH & Co. KG, 2012 (cit. on pp. 2, 4, 10–14, 16–20, 22–26, 28, 29, 38–48, 50, 52–55, 57–62, 64, 66, 68, 70, 71, 95, 98).

[Hea06] Jeff Heaton. *Programming spiders, bots, and aggregators in Java*. John Wiley & Sons, 2006 (cit. on p. 19).

[Hsi14] Cheng Hsiao. *Analysis of panel data*. 54. Cambridge university press, 2014 (cit. on p. 35).

[Hud+18] Miftachul Huda et al. "Big data emerging technology: insights into innovative environment for online learning resources." In: *International Journal of Emerging Technologies in Learning (iJET)* 13.1 (2018), pp. 23–36 (cit. on p. 8).

[HFC11] Bradley Huffaker, Marina Fomenkov, and K Claffy. "Geocompare: a comparison of public and commercial geolocation databases." In: *Proc. NMMC* (2011), pp. 1–12 (cit. on pp. 48, 49).

[IBM11] IBM. *The 2011 IBM Tech Trends Report – Tech Trends of today. Skills for tomorrow.* 2011 (cit. on p. 10).

[Jan09] Bernard J Jansen. "Understanding user-web interactions via web analytics." In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1.1 (2009), pp. 1–102 (cit. on pp. 2, 4, 7, 13, 14, 18–24, 26, 27, 29, 36, 37, 50–54, 56–58).

[JS05] David Janzen and Hossein Saiedian. "Test-driven development concepts, taxonomy, and future direction." In: *Computer* 38.9 (2005), pp. 43–50 (cit. on p. 109).

[JK15] Joel Järvinen and Heikki Karjaluoto. "The use of Web analytics for digital marketing performance measurement." In: *Industrial Marketing Management* 50 (2015), pp. 117–127 (cit. on pp. 37, 40).

[Kai+13] Stephen Kaisler et al. "Big data: Issues and challenges moving forward." In: *2013 46th Hawaii international conference on System sciences (HICSS)*. IEEE. 2013, pp. 995–1004 (cit. on p. 9).

[KH10] Andreas M Kaplan and Michael Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media." In: *Business horizons* 53.1 (2010), pp. 59–68 (cit. on p. 70).

[KWG13] Avita Katal, Mohammad Wazid, and RH Goudar. "Big data: issues, challenges, tools and good practices." In: *Contemporary Computing (IC3), 2013 Sixth International Conference on*. IEEE. 2013, pp. 404–409 (cit. on p. 9).

[Kau10] Avinash Kaushik. *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity*. Indianapolis: John Wiley & Sons, 2010 (cit. on pp. 2–4, 11–13, 16, 17, 22, 31, 34–37, 39, 41–43, 45, 46, 50–53, 55–59, 61, 63, 64, 66, 68, 98).

[Kau07] Avinash Kaushik. *Web analytics: An hour a day*. John Wiley & Sons, 2007 (cit. on pp. 2, 3, 10–20, 22–24, 26–32, 34–39, 41, 44, 45, 53–56).

[KBR15] Stefan Keil, Peter Böhm, and Marc Rittberger. "Qualitative Web Analytics: New Insights into Navigation Analysis and User Behavior - A Case Study of the German Education Server." In: May 2015 (cit. on pp. 4, 11, 31, 34, 35).

[Ken+11] Michael L Kent et al. "Learning web analytics: A tool for strategic communication." In: *Public Relations Review* 37.5 (2011), pp. 536–543 (cit. on pp. 46, 47, 53, 58).

[KRS02] Ron Kohavi, Neal J Rothleder, and Evangelos Simoudis. "Emerging trends in business analytics." In: *Communications of the ACM* 45.8 (2002), pp. 45–48 (cit. on p. 2).

[Koh+04] Ron Kohavi et al. "Lessons and challenges from mining retail e-commerce data." In: *Machine Learning* 57.1 (2004), pp. 83–113 (cit. on p. 19).

[KR14] Krish Krishnan and Shawn P Rogers. *Social data analytics: Collaboration for the enterprise*. Newnes, 2014 (cit. on p. 49).

[KSK12] Lakhwinder Kumar, Hardeep Singh, and Ramandeep Kaur. "Web analytics and metrics: a survey." In: *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*. ACM. 2012, pp. 966–971 (cit. on pp. 18, 24–26, 29–32, 37, 98).

[Lan01] Doug Laney. "3D data management: Controlling data volume, velocity and variety." In: *META Group Research Note* 6 (2001), p. 70 (cit. on pp. 8, 9).

[Law12] K Blaine Lawlor. "Smart goals: How the application of smart goals can contribute to achievement of student learning outcomes." In: *Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference*. Vol. 39. 2012 (cit. on p. 14).

[Lei+09] Barry M Leiner et al. "A brief history of the Internet." In: *ACM SIGCOMM Computer Communication Review* 39.5 (2009), pp. 22–31 (cit. on p. 6).

[LV16] Katherine N Lemon and Peter C Verhoef. "Understanding customer experience throughout the customer journey." In: *Journal of Marketing* 80.6 (2016), pp. 69–96 (cit. on pp. 111–114, 118, 123, 133).

[Lew13] Anna Lewis. *Search Engine Watch – How to Use the Google Analytics Frequency & Recency Report*. July 2013. URL: `https://searchenginewatch.com/sew/how-to/2282540/how-to-use-the-google-analytics-frequency-recency-report` (visited on 05/10/2019) (cit. on pp. 46, 47).

[LCC13] Ee-Peng Lim, Hsinchun Chen, and Guoqing Chen. "Business intelligence and analytics: Research directions." In: *ACM Transactions on Management Information Systems (TMIS)* 3.4 (2013), p. 17 (cit. on pp. 10–12).

[LL05] Sonia Livingstone and Leah A Lievrouw. *Handbook of new media*. Sage Publications, 2005 (cit. on p. 6).

[Lof12] Wayne Loftus. "Demonstrating success: Web analytics and continuous improvement." In: *Journal of Web Librarianship* 6.1 (2012), pp. 45–55 (cit. on p. 13).

[Low] Lisa Lowe. *Social Pilot – 125 Amazing Social Media Statistics You Should Know in 2016*. URL: `https://web.archive.org/web/20161024124411/https://socialpilot.co/blog/125-amazing-social-media-statistics-know-2016/` (visited on 05/10/2019) (cit. on p. 8).

[Mar11] Kate Marek. "Installing and Configuring Google Analytics." In: *Library Technology Reports* 47.5 (2011), pp. 17–25 (cit. on pp. 25, 26).

[Mar17] Bernard Marr. *Forbes – Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud*. Jan. 2017. URL: https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/ (visited on 07/18/2017) (cit. on p. 8).

[Max] MaxMind. *MaxMind – GeoIP2 City Accuracy*. URL: https://www.maxmind.com/en/geoip2-city-database-accuracy (visited on 05/10/2019) (cit. on p. 49).

[MM12] Jonathan R Mayer and John C Mitchell. "Third-party web tracking: Policy and technology." In: *2012 IEEE Symposium on Security and Privacy*. IEEE. 2012, pp. 413–427 (cit. on pp. 32, 35).

[MB+12] Andrew McAfee, Erik Brynjolfsson, et al. "Big data: the management revolution." In: *Harvard business review* 90.10 (2012), pp. 60–68 (cit. on pp. 8, 9).

[McF05] Christopher McFadden. "Optimizing the online business channel with web analytics." In: *Cited on pp. xi and* 45 (2005) (cit. on pp. 6, 7, 11–16, 49).

[Mer14] David Mercer. *Strategy Analytics – 33 Billion Internet Devices By 2020: Four Connected Devices For Every Person In World*. Oct. 2014. URL: www4.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=5609 (visited on 07/18/2017) (cit. on p. 7).

[Mey08] Dr. Peter J. Meyers. *Converting The Believers - How to Turn Website Visitors into Buyers*. Chicago: User Effect, 2008 (cit. on p. 14).

[Mon+04] Alan L Montgomery et al. "Modeling online browsing and path analysis using clickstream data." In: *Marketing science* 23.4 (2004), pp. 579–595 (cit. on pp. 11, 22).

[Naj05] Lawrence J Najjar. "Designing E-commerce User Interfaces." In: *Handbook of Human Factors in Web Design*. 2005, pp. 584–595 (cit. on p. 65).

[NC11] Kazuo Nakatani and Ta-Tao Chuang. "A web analytics tool selection method: an analytical hierarchy process approach." In: *Internet Research* 21.2 (2011), pp. 171–186 (cit. on pp. 11, 17).

[Neg04] Solomon Negash. “Business intelligence.” In: *Communications of the association for information systems* 13.1 (2004), p. 15 (cit. on p. 10).

[Net18] Mozilla Developer Network. *MDN – Async scripts for asm.js*. July 2018. URL: `https://developer.mozilla.org/en-US/docs/Games/Techniques/Async_scripts` (visited on 05/10/2019) (cit. on p. 103).

[Net19a] Mozilla Developer Network. *MDN – IIFE*. Mar. 2019. URL: `https://developer.mozilla.org/en-US/docs/Glossary/IIFE` (visited on 05/10/2019) (cit. on p. 105).

[Net19b] Mozilla Developer Network. *MDN – script - The Script element*. Mar. 2019. URL: `https://developer.mozilla.org/en-US/docs/Web/HTML/Element/script` (visited on 05/10/2019) (cit. on pp. 103, 105).

[Net19c] Mozilla Developer Network. *MDN – Structuring the Web*. May 2019. URL: `https://developer.mozilla.org/en-US/docs/Learn/HTML/Introduction_to_HTML/Getting_started` (visited on 05/10/2019) (cit. on p. 100).

[NH05] Tom Noda and Shawn Helwig. “Rich internet applications.” In: *Technical Comparison and Case Studies of AJAX, Flash, and Java based RIA UW E-Business-Consortium Opinion Papers* (2005) (cit. on pp. 4, 6, 7, 10, 17, 19, 22, 26, 66).

[NZS06] Jean-Pierre Norguet, Esteban Zimányi, and Ralf Steinberger. “Improving web sites with web usage mining, web content mining, and semantic analysis.” In: *SOFSEM*. Springer. 2006, pp. 430–439 (cit. on p. 10).

[Ogl10] James A Ogle. “Improving Web Site Performance Using Commercially Available Analytical Tools.” In: *Clinical Orthopaedics and Related Research®* 468.10 (2010), pp. 2604–2611 (cit. on pp. 2, 11–13, 19, 37, 55).

[PPC12] Heikki Pakkala, Karl Presser, and Tue Christensen. “Using Google Analytics to measure visitor statistics: The case of food composition websites.” In: *International Journal of Information*

Management 32.6 (2012), pp. 504–512 (cit. on pp. 12, 32, 38, 43, 52, 53, 58, 63).

[PC10] Rajni Pamnani and Pramila Chawan. "Web Usage Mining: A research area in Web mining." In: *Proceedings of ISCET* (2010), pp. 73–77 (cit. on pp. 9, 17, 19, 20, 22).

[Pan+11] Saroj K Pani et al. "Web usage mining: A survey on pattern extraction from web logs." In: *International Journal of Instrumentation, Control & Automation* 1.1 (2011), pp. 15–23 (cit. on pp. 7, 11, 16–20, 23, 26, 31).

[PH06] Sophia Parker and Joe Heapy. "The journey to the interface." In: *London: Demos* (2006) (cit. on p. 111).

[Par12] Article 29 Data Protection Working Party. *Article 29 – Opinion 04/2012 on Cookie Consent Exemption*. Apr. 2012. URL: `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2012/wp194_en.pdf` (visited on 10/29/2018) (cit. on p. 95).

[Pea02] Darren Peacock. "Statistics, Structures & Satisfied Customers: Using Web Log Data to Improve Site Performance." In: (2002) (cit. on pp. 37, 38).

[Pir07] VH Pirttimaki. "Conceptual analysis of business intelligence." In: *South African journal of information management* 9.2 (2007), pp. 1–1 (cit. on p. 10).

[Pra+13] Adhi Prasetio et al. "The Impact of Traffic Source on Page Views." In: *Proceedings of the 7th International Conference on Information & Communication Technology and Systems*. 2013, pp. 15–16 (cit. on p. 58).

[PF13] Foster Provost and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." In: *Big data* 1.1 (2013), pp. 51–59 (cit. on p. 11).

[RGS09] Arvind Rangaswamy, C Lee Giles, and Silvija Seres. "A strategic perspective on search engines: Thought candies for practitioners and researchers." In: *Journal of Interactive Marketing* 23.1 (2009), pp. 49–60 (cit. on p. 36).

[Ran09] Jayanthi Ranjan. "Business intelligence: Concepts, components, techniques and benefits." In: *Journal of Theoretical and Applied Information Technology* 9.1 (2009), pp. 60–70 (cit. on p. 10).

[RDR07] Matthew Richardson, Ewa Dominowska, and Robert Ragno. "Predicting clicks: estimating the click-through rate for new ads." In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, pp. 521–530 (cit. on p. 37).

[ROR17] Mark S Rosenbaum, Mauricio Losada Otalora, and German Contreras Ramirez. "How to create a realistic customer journey map." In: *Business Horizons* 60.1 (2017), pp. 143–150 (cit. on p. 112).

[San87] Robert Sanders. "The Pareto principle: its use and abuse." In: *Journal of Services Marketing* 1.2 (1987), pp. 37–40 (cit. on pp. 59, 61).

[SA08] John Seely Brown and RP Adler. "Open education, the long tail, and learning 2.0." In: *Educause review* 43.1 (2008), pp. 16–20 (cit. on pp. 59, 61).

[Ser13] Government Digital Service. *Government Digital Service – How many people are missing out on JavaScript enhancement?* Oct. 2013. URL: `https://gds.blog.gov.uk/2013/10/21/how-many-people-are-missing-out-on-javascript-enhancement/` (visited on 05/10/2019) (cit. on p. 27).

[Sha12] Sayf Sharif. *bounteous – Where Should The Google Analytics Tracking Code Be Placed*. Feb. 2012. URL: `https://www.bounteous.com/insights/2012/02/09/where-should-google-analytics-tracking-code-be-placed/?ns=l` (visited on 05/10/2019) (cit. on p. 26).

[Sha19a] Himanshu Sharma. *Optimize smart – Virtual Pageviews in Google Analytics*. May 2019. URL: `https://www.optimizesmart.com/event-tracking-guide-google-analytics-simplified-version/` (visited on 05/10/2019) (cit. on pp. 68, 70).

[Sha19b] Himanshu Sharma. *Optimize smart – Virtual Pageviews in Google Analytics - Complete Guide*. May 2019. URL: `https://www.optimizesmart.com/virtual-pageviews-google-analytics-complete-guide/` (visited on 05/10/2019) (cit. on pp. 66–68).

[SKS14] Himani Singal, Shruti Kohli, and Amit Kumar Sharma. "Web analytics: State-of-art & literature assessment." In: *2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence)*. IEEE. 2014, pp. 24–29 (cit. on pp. 29, 37).

[Sla12] Wolfgang Slany. "A mobile visual programming system for Android smartphones and tablets." In: *Visual Languages and Human-Centric Computing (VL/HCC), 2012 IEEE Symposium on*. IEEE. 2012, pp. 265–266 (cit. on p. 76).

[Sla14] Wolfgang Slany. "Tinkering with Pocket Code, a Scratch-like programming app for your smartphone." In: *Proc. of Constructionism* (2014) (cit. on pp. 2, 73, 75).

[Sta18] Statista. *Statista – Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)*. July 2018. URL: `https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/` (visited on 07/18/2018) (cit. on p. 7).

[SZ09] Marc Stickdorn and Anita Zehrer. "Service design in tourism: Customer experience driven destination management." In: *First Nordic conference on service design and service innovation, Oslo*. 2009, pp. 1–16 (cit. on pp. 110–113, 118, 123, 133).

[Sul+14] Snezhana Sulova et al. "Evaluation of E-commerce Web sites on the Basis of Usability Data." In: *Izvestiya* 3 (2014), pp. 37–47 (cit. on pp. 12, 31).

[SK09] KR Suneetha and Raghuraman Krishnamoorthi. "Identifying user behavior by analyzing web server access log file." In: *IJCSNS International Journal of Computer Science and Network Security* 9.4 (2009), pp. 327–332 (cit. on pp. 17–19, 22, 26).

[Tan12] Kenji Tanaka. "NPO Marketing for Project 'Catrobat'." In: Presentation, 2012 (cit. on pp. 76, 77).

[Tem+10] Bruce D Temkin et al. "Mapping the customer journey." In: *Forrester Research* (2010), p. 3 (cit. on p. 112).

[The09] Michael Thelwall. *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan and Claypool Publishers, 2009 (cit. on p. 16).

[Tul+02] Tom Tullis et al. "An empirical comparison of lab and remote usability testing of web sites." In: *Usability Professionals Association Conference*. 2002 (cit. on p. 15).

[W19] Rob W. *Mozilla Developer Network — Set-Cookie*. Mar. 2019. URL: `https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Set-Cookie` (visited on 05/10/2019) (cit. on pp. 31, 32).

[WK09] Daniel Waisberg and Avinash Kaushik. "Web Analytics 2.0: empowering customer centricity." In: *The original Search Engine Marketing Journal* 2.1 (2009), pp. 5–11 (cit. on pp. 27, 29, 30).

[Web13a] Jonathan Weber. *koozai – The Power of Segmentation in Web Analytics*. Oct. 2013. URL: `https://www.koozai.com/blog/analytics/measurefest-power-segmentation-web-analytics/` (visited on 05/10/2019) (cit. on pp. 38–40, 49, 50).

[Web13b] Jonathan Weber. *Lunametrics – The Locations Report in Google Analytics*. Mar. 2013. URL: `https://www.bounteous.com/insights/2013/03/19/locations-report-google-analytics/?ns=1` (visited on 05/10/2019) (cit. on p. 49).

[Wee13] Electronics Weekly. *Fifty billion internet nodes predicted by 2020*. Jan. 2013. URL: `https://www.electronicsweekly.com/news/business/information-technology/fifty-billion-internet-nodes-predicted-by-2020-2013-01/` (visited on 05/10/2019) (cit. on p. 6).

[WH06] Birgit Weischedel and Eelko KRE Huizingh. "Website optimization with web metrics: a case study." In: *Proceedings of the 8th international conference on Electronic commerce: The new e-commerce: innovations for conquering current barriers, obstacles* *and limitations to conducting successful business on the internet*. ACM. 2006, pp. 463–470 (cit. on pp. 12, 20, 22).

[Wu+09] Jingxuan Wu et al. "Using web-analytics to optimize education website." In: *International Conference on Hybrid Learning and Education*. Springer. 2009, pp. 163–174 (cit. on pp. 10, 11).

[Xue04] Susan Xue. "Web usage statistics and Web site evaluation: a case study of a government publications library Web site." In: *Online Information Review* 28.3 (2004), pp. 180–190 (cit. on pp. 37, 38).

[Yea01] Jane Yeadon. "Web site statistics." In: *Vine* 31.3 (2001), pp. 55–60 (cit. on pp. 37, 38).

[Zak10] Nicholas C. Zakas. *YAHOO! Developer Network — How many users have JavaScript disabled?* Oct. 2010. URL: `http://web.archive.org/web/20110218221930/http://developer.yahoo.com/blogs/ydn/posts/2010/10/how-many-users-have-javascript-disabled/` (visited on 12/21/2017) (cit. on p. 27).

[Zec14] Ashley Zeckman. *Search Engine Watch – Organic Search Accounts for Up to 64% of Website Traffic*. July 2014. URL: `https://searchenginewatch.com/sew/study/2355020/organic-search-accounts-for-up-to-64-of-website-traffic-study` (visited on 05/10/2019) (cit. on pp. 56, 61, 62, 119).

[ZP15] Guangzhi Zheng and Svetlana Peltsverger. "Web Analytics Overview." In: *Encyclopedia of Information Science and Technology, Third Edition*. IGI Global, 2015, pp. 7674–7683 (cit. on pp. 4, 29, 34).

[Zil15] Jakub Zilincan. "Search Engine Optimization." In: *CBU International Conference Proceedings*. Vol. 3. 2015, pp. 506–510 (cit. on pp. 42, 62, 64).