Analyzing and Visualizing Usage for Common Coordinate Framework User Interfaces Using Google Analytics Data

HuBMAP analytics

Mustafa Alsaegh, Xuemei Hu, Arpan Ojha, Amanda Turney
Indiana University, Bloomington, USA
malsaegh@iu.edu, xh18@iu.edu, arojha@iu.edu, turneya@iu.edu

I. Introduction

Google Analytics is an important tool to track and report website traffic and is now the most widely used web analytics service on the internet. Our client utilizes Google Analytics to log the usage of their common coordinate framework [1] that is Registration User Interface(RUI) [2], Exploration User Interface(EUI) [3] and ASCT+B Reporter [4]. In an effort to improve these three user interfaces, our client recently implemented usage tracking via Google Analytics. Specifically, they used custom events to log mouse movement, clicks and organ tables loaded. A statistical study can provide insight into the different ways users are interacting with the ASCT+B Reporter visualization with a particular focus on the insight needs of distributions and comparison. Our paper performs analysis to identify different user experiences in the HuBMAP platform.

II. Visualization Goals and Insight Needs

A. User Event Frequencies

The overarching goal of the client is to understand how users are interacting with the ASCT+B Reporter application in order to gain insight of where and how to improve the user interface. One research interest of the client was to understand the distribution of user events such as page views, clicks, and scrolls.

B. Organ Tables

The goal is to understand oval frequency. A simple trend bar graph on the number of searches divided across 24 hours of the day.

C. Data Export Formats

The goal is to understand what types of export formats are being used by the users. We analyzed the organ frequencies for the exports made by the users of the platform.

D. Comparison Between Same Time Periods Across Different Years

The goal is to understand the major differences from the same time frame between 2021-2022 and 2022-2023.

III. Data Acquisition

The datasets were given access from the same time frame in 20211210 - 20220212 and 20221210 - 20230212. And the data comes as a CSV file.
During 20221210-20230212, the datasets include 17529 user pseudo ids. During 20211210-20220212, the datasets include 39599 user pseudo ids. The users were reduced half this year compared to last year. Both datasets include 9 columns which are 1. User_pseudo_id 2. Hostname 3. Page_location 4. Event_date 5. User first touch Timestamp 6. Event_timestamp 7.Event_name 8. Event_category 9. Event_label.

IV. Discussion of related work

Due to the popularity of Google Analytics data, there are many different dashboards and visualizations examples that can be found online [5]. Page view events are tracked over time to visualize the temporal component of the usage data as well as aggregated by count over the different pages within the application, which is similar to the prototype visualization in the reference. Another example of Google Analytics data visualization is the Sankey Graph showing the temporal flow of how users navigate through the CNS webpage as illustrated in the Atlas of Knowledge on page 49 [6]. Lastly, this Tableau dashboard [7] visualizes website clicks per hour as well as a heatmap overlaid on an image of the web page to visualize where the webpage is getting clicked.

V. Simple Statistics

Based on the user _psuedo_id column in the dataset, there are 552 unique users that have visited anywhere in the website application. However, when looking only at the visualization tool webpage specifically, there are 314 unique users. Over 17000 page visits originated from a total of 110 unique users for the most recent time period of Google Analytics records (December 2022-February 2023). Most of these users were very evenly distributed in regards to their purpose. There are a total of 28 organs. Small intestines were by far the most frequently searched organ with 1601 searches followed by 1594 lung searches. Six organs were searched over 1000 times which were small intestine, lungs, blood, kidney, liver and heart. We had some missing data which amounted to 10

Identify applicable funding agency here. If none, delete this.

percent of data loss. However we didn't have any losses in organ searches. Three unique hostnames were visited. Finally, the mousemove event category was 2500 times more frequent than the second most used in that category. This suggests that people spend quite some time trying to understand the Hubmap. The analysis conducted on the data revealed some interesting insights about the export activities of the users. Although the original files had more than 17,000 rows each, filtering the event label column to include only the export events reduced the data to a mere 96 rows. One notable finding is that PNG is the most popular file format for exports, with a count of 25. We also determined that out of the 110 users on the platform, 28 had exported data at least once. This represents approximately 25 percent of the total users. This small dataset provided valuable insights into the most frequently exported file formats, the number of users, and what organ tables were exported. However, it is important to note that these findings are based on a limited dataset, and may not be representative of the larger user population. Further analysis using a larger sample size or additional data sources could provide a more accurate picture of user behavior and preferences when it comes to exporting data.

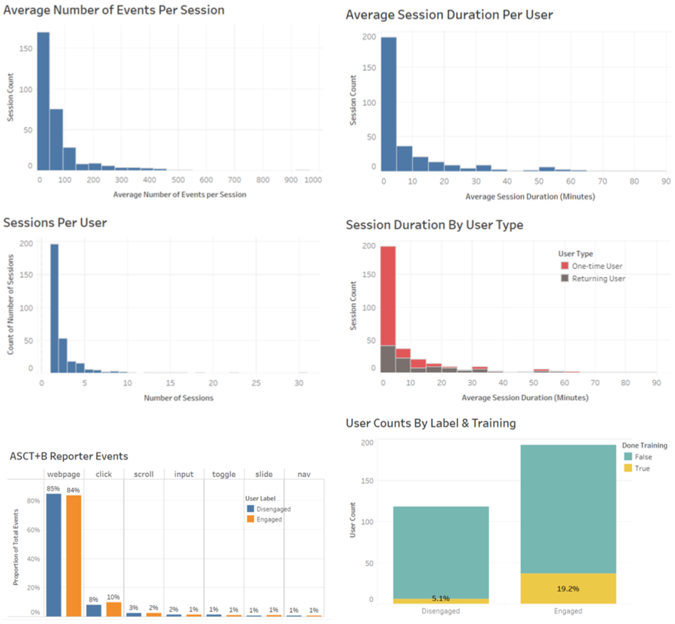

Fig. 1. The distribution of **top left**. number of session visits per user.**top right**, average session duration per user. **middle left**, the average number of events per session. **middle right**. Same figure as figure 1 but grouped by returning versus one-time users. **bottom left**. Distribution of frequency of event types grouped by engagement user status.**bottom right**. Stacked bar chart of counts of trained or un-trained users across the disengaged versus engaged user status.

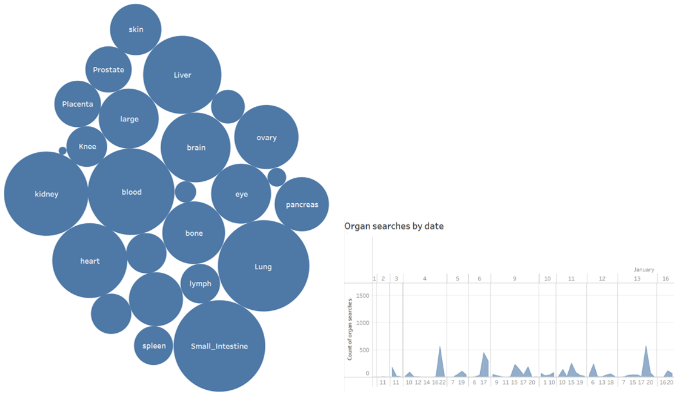

Fig. 2. **left** Bubble chart of organs and their search frequency. **right** Hourly organ search rates over time.

VI. Data Analysis & Visualizations

A. User Event Frequencies

Feature engineering was performed on the Google Analytics logs to capture important characteristics of how users interact with the web application. Utilizing the "session_start" event type log, user sessions were labelled, and individual event logs were tied to a session ID. From there, features such as the user session duration and number of events per session were calculated. Lastly, while the user_pseudo_id does not have the capability to track a user across multiple clients/devices, it can track a user over time on a single device [8]. With assumed very low cross-platform use (desktop and mobile) due to the nature of the application, user _pseudo_ids are a reasonable representation of the users. Therefore, the number of sessions per user, average session duration per user, and average number of events per user were defined. A flag was also created to determine whether a user was disengaged based on the criteria of a single session that had a duration of 2 minutes or less. Lastly, the application has a plethora of resources on how to use the visualization tool available in documentation, tutorials, and videos. Given the user event logs, it can be seen whether a user has visited any of those resources or not, and thus a flag for whether that user had consulted any training resources or not was also captured.
The analysis of user events is primarily a statistical analysis and resulted in a series of statistical visualizations to meet the insight needs of distribution and comparison.

B. Organ Tables

Topical, temporal, and network visualizations were used to gain insights of the organ tables accessed by users. The topical analysis was accomplished through a bubble chart which shows the most frequently searched organ. The temporal visualization is a bar graph to show the time of day when the organs were searched the most. Finally, a network tree visualization was developed to map the most frequently correlated organ searches by user. The organs are represented by nodes and the edges signify connections in an undirected graph. The edge weight represents the frequency of the connection and was built using python's pyviz.

C. Data Export Formats

Data exports were grouped by the organ tables for the data being exported and a bar graph showing the frequency of export by organ is shown in figure 4. The frequency of exports based on the format type (PNG, SVG, OWL, etc.) were also analyzed and a bar graph showing frequency for each format type is shown in figure 4. Lastly, an analysis of the number

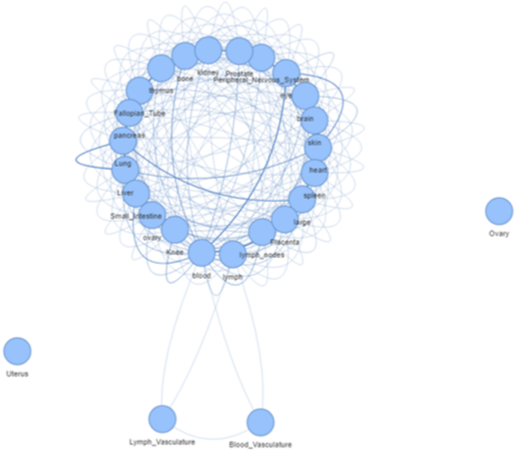

Fig. 3. Network visualization showing organs searched together. Nodes represent organs and edges represent organ pairs searched together with edge width representing the frequency.

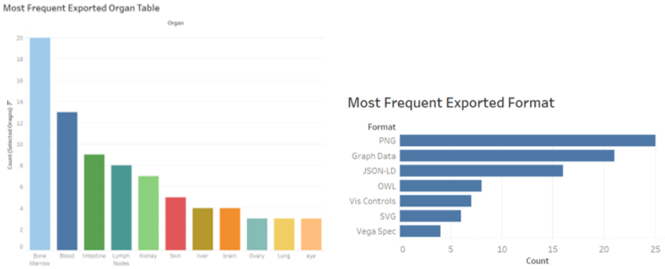

Fig. 4. **left** Frequency of exports by format type. **right** Frequency of exports by format type.

of exports by each user was done and the top 5 users with the most exports are shown in figure 5.

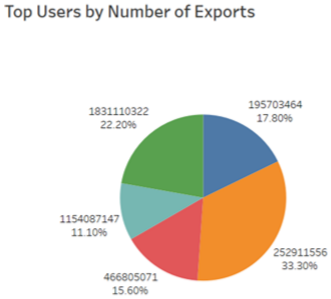

Fig. 5. Top 5 users by number of exports.

D. *Comparison Between Same Time Periods Across Different Years*

One of the research questions of interest was to look for differences in the Google Analytics logs between the same time period across different years. A temporal trend of weekly record counts can be compared between the different years. Figure 6 shows this comparison between the December 2021 to February 2022 time period and the December 2022 to February 2023 time period. The distribution of user event frequencies can also be compared across the years. Figure 6 shows the counts of the different user events for each time period in a stacked bar chart.

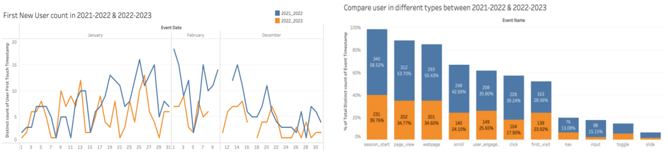

Fig. 6. **left** Line plot showing the weekly new user counts between the same time period across two different subsequent years. **right** Stacked bar chart with the user event frequencies for the time period across the two different years (blue is 2021-2022 and orange is 2022 to 2023).

VII. Discussion of Key Insights

A. *User Event Frequencies*

Overall, we see that the number of sessions per user, the average session duration per user, and the average counts of events per session all have very right skewed distributions. This means that a lot of traffic on the visualization tool web page is coming from one-time users and/or very short sessions. We see that one-time users actually contribute the most to the number of very short session durations. However, there were some one-time users that still spent a substantial amount of time using the visualization tool. When comparing the user event frequency between the engaged versus disengaged users, it doesn't look like there is much of a difference in the type of events coming from two types of users. However, when we look at the training rates of the engaged versus disengaged users, we see that the engaged users have a 3x rate of having viewed any sort of training/tutorial documentation or video within the application. This could be an indicator that some of one-time and short visits could be due to users not fully understanding how to use the visualization tool.

B. *Organ Tables*

The topical bubble chart suggests that the small intestine is the organ most searched within the tool most closely followed by the lung. This is important because the ASCT+B app developers might reach out to the researchers working in gastrointestinal/metabolic or respiratory space in order to get more specific user feedback. The temporal trends show that the most frequent searches are made between 1 and 4pm EST suggesting usage after lunch or before leaving for work. However, the application has users across the country and globe so this might just suggest that most users are probably located within the United States and thus using it during US daytime hours.

The network visualization shows that the uterus and ovary organs searches are never paired with any other organs. Lymph vasculature and blood vasculature are also very infrequently searched with any other organs. However, the remaining organs have a densely connected map with lymph, spleen, and skin being some of the most frequently paired searches followed by blood and eye.

C. Data Export Formats

The analysis of date export formats identified several key findings. First, it was found that out of 110 users, 28 had exported data at least once, representing 25% of the total users suggesting this is a commonly used feature of the application. Secondly, the PNG format was the most exported format type followed by the Graph Data and JSON-LD. Thirdly, the examination of the frequency of exports by organ show that bone marrow was the most frequently exported organ followed by blood, lymph node, and kidney. Lastly, the highest number of exports occurred on December 14th, 2021.

D. Comparison Between Same Time Periods Across Different Years

In summary, the users had reduced by half for the most recent time period (2022-2023) compared to the year prior. Once of the reasons is due to overall usage being spread between the three user interfaces developed between those time periods: the RUI, EUI, and the ASCT+B Reporter. Another reason was increased usage of the application more than the website in the later year.

There is also a difference in the weekly count of records between the two years shown in figure 6. Both time periods show the weekly count increasing in the second week of December but then dropping right before Christmas. However, the weekly counts continue to increase after the new year in 2023 whereas it decreases in 2022. Figure 6 shows that there was not much of a difference in the distribution of user event frequencies from the earlier time period to the more recent time period.

VIII. Problems Analysis

Validation is a critical aspect of any research project, but it can also be a challenging and time-consuming process. In this section, we discuss some of the problems we encountered during the validation phase of our project and how we solved them. One of the major issues we faced was the need to clean and preprocess the data before analysis. We found that the ”event_label” column contained additional information that needed to be filtered out using Python, which added an extra layer of complexity to our analysis. To address this issue, we implemented better data cleaning and preprocessing techniques, which allowed us to work with a more accurate and reliable dataset.

Another issue we encountered was the need to compare data from the same time period across different years. We found that the data set decreased by half in the 2022-2023 period compared to the previous year, which made it difficult to draw firm conclusions. To address this issue, we disentangled events by page_location, which helped us to determine whether the user was a developer or a legitimate user. This approach allowed us to figure out all the events and draw more accurate conclusions.

IX. Challenges and Opportunities

During the User Event Frequencies project, we faced several challenges, including identifying usage patterns of true users and understanding how Google Analytics creates pseudo_user_ids. In the future, we could improve this analysis by separating developers and testers into a separate environment and conducting user interviews to augment our analysis. Additionally, we identified unfamiliarity with new tools and data cleaning as significant restrictions in the Organ Tables project. Further data columns and context of the dataset would benefit future analysis.

In the Comparison Between Same Time Periods Across Different Years project, the limited dataset presented a significant challenge in drawing firm conclusions and identifying meaningful trends. However, future work could include expanding the scope of the analysis to include a larger dataset, exploring additional variables such as user demographics and feedback, and developing more advanced analytical models to identify patterns and trends. Overall, these opportunities could yield more comprehensive and robust analysis, driving further improvements to the platform.

Despite the challenges encountered during the Data Export Formats project, we implemented improvements to the data collection process and introduced new features to improve the user experience. However, the limited amount of data available for analysis and uneven distribution of export activity among users presented further challenges. To address these challenges, future work could include expanding the scope of the analysis to include a larger dataset, exploring additional variables and features, and developing more advanced analytical models to identify patterns and trends. These opportunities could lead to more meaningful insights and improvements to the platform.

X. Acknowledgements

The authors would like to thank Abbay Bharat Rajde to support as Stakeholder in this project. In addition, to the classmates who provided valuable feedback to our team. Also thanks a lot to Dr. Andreas Buckle, Michael Ginda and all the Teaching Assistants in the E583 Information Visualization Course. Special thanks to Dr. Katy Börner for the wonderful lecture.

References

[1] HuBMAP: CCF Portal https://hubmapconsortium.github.io/ccf/

[2] Registration User Interface https://hubmapconsortium.github.io/ccf-ui/rui/

[3] Exploration User Interface https://portal.hubmapconsortium.org/ccf-eui

[4] ASCT+B Repoter https://hubmapconsortium.github.io/ccf-asct-reporter/

[5] Tableau dashboard example https://public.tableau.com/app/profile/technical.product.marketing/viz/WebsiteDashboardDemo_10_0/TrafficTrends

[6] Börner, Katy. 2015. Atlas of Knowledge. The MIT Press. https://scimaps.org/books

[7] Tableau dashboard example https://public.tableau.com/app/profile/rachel7046/viz/HeatMapofMainPage_15602656355340/WebsiteClicks

[8] What is user_pseudo_id in GA4 BigQuery Export? https://www.optizent.com/blog/what-is-user_pseudo_id-in-ga4-bigquery-export/