Exploration on the Use of WDQS

Breakdown by Geography, User Agent and Referer Class

Chelsy Xie (Analysis & Report)
Mikhail Popov (Review)
Deb Tankersley (Review)
Stas Malyshev (Review)

19 September 2016

Executive Summary

Wikidata Query Service (WDQS) was launched publicly on September 7, 2015. For the first anniversary, we want to take a look into who is using WDQS, and how they are using it. In this report, we focus on the web requests to the SPARQL endpoint, their breakdown by country, user agent, referer class, and their pattern over time.

We found that Germany, United States and France have the largest number of users and queries. Among all operating systems, Mac OS X has the largest number of real users and Ubuntu users submit the most queries; among all browsers, Chrome is the most popular one and Firefox users submit the most queries. Most queries do not have a referer as they are from real people, but there are a large amount of queries that are from automata. There are weekly cycles in the number of queries and users. We also saw an increasing trend in the number of users.

Data

Extracting successful (HTTP status codes 200 & 304) web requests to the SPARQL endpoint from July 1st to August 29, 2016, we count the number of queries and users by country, user agent and referer class. Here the "user" is identified by the combination of client IP and user agent, since different devices, OS's, and browsers on the same network may share the same IP address. However, if users update their browsers and OS's to a newer version in the middle of the day, they would be counted as two users that day. See data.R for more details.

Results

Cross-Sectional

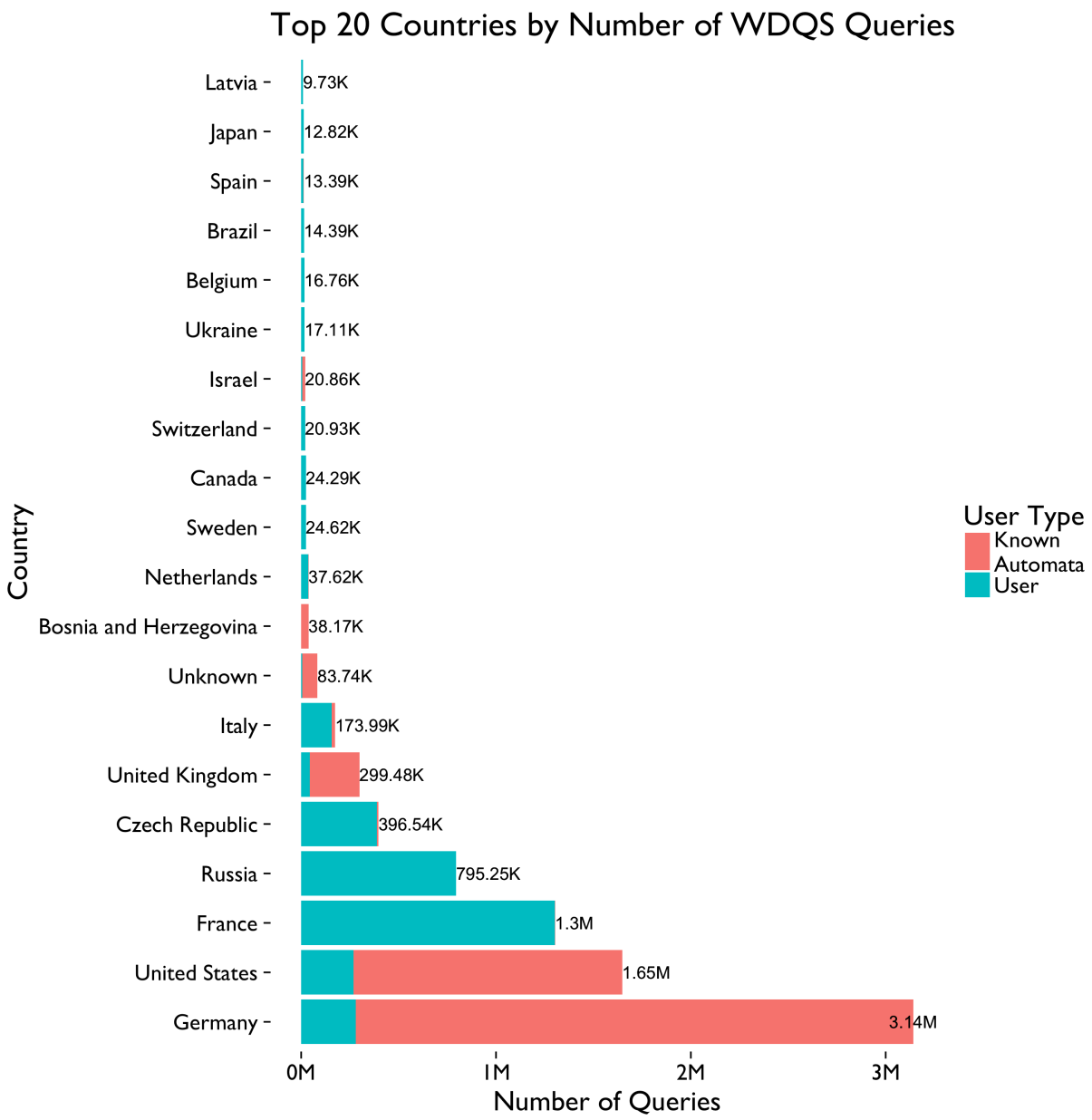

Figure 1: Germany, United States and France take the first 3 places on the rank. While most of them are automata queries in Germany and US, France has the most user queries.

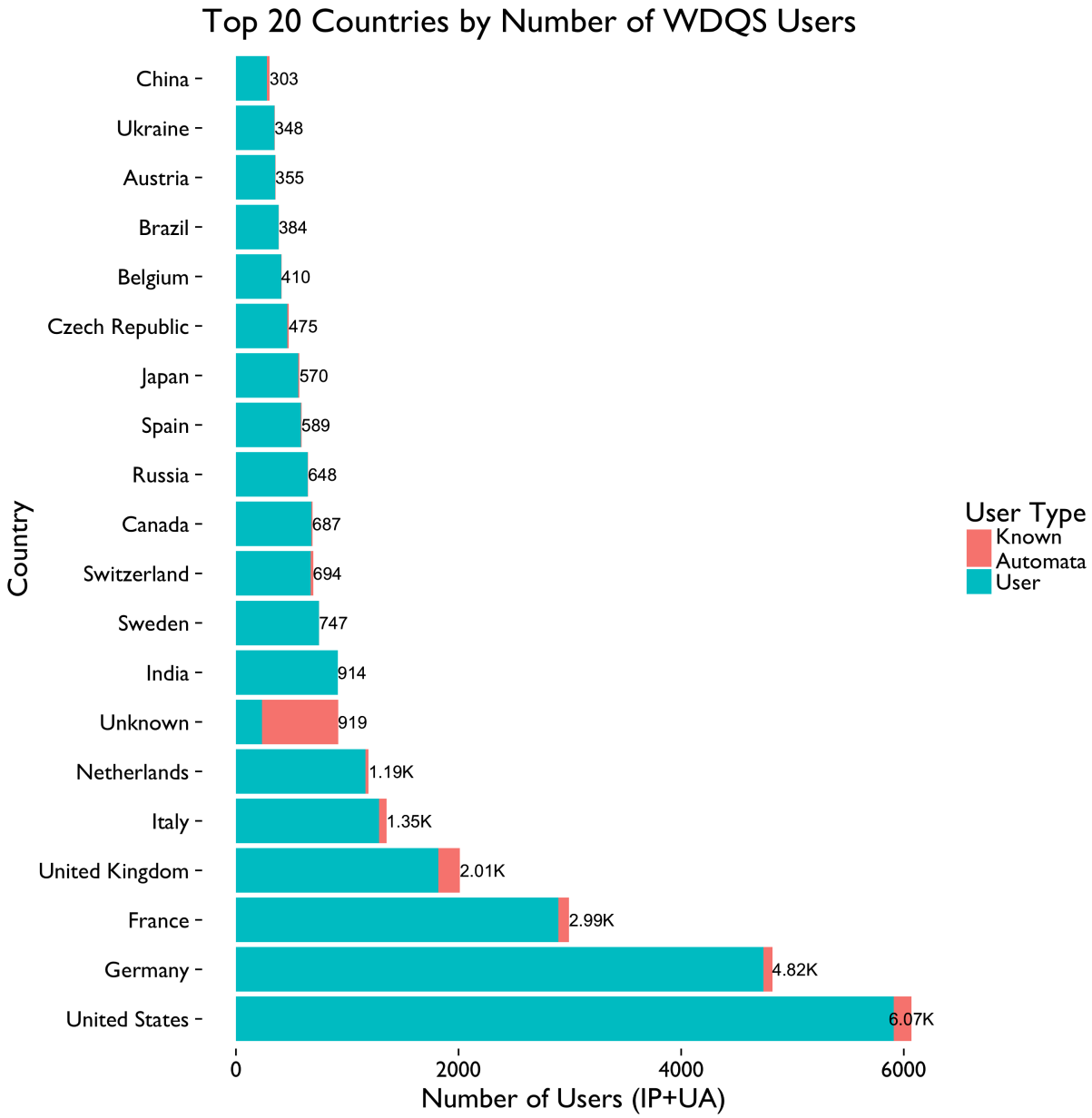

Figure 2: United States, Germany and France have the largest number of WDQS users.

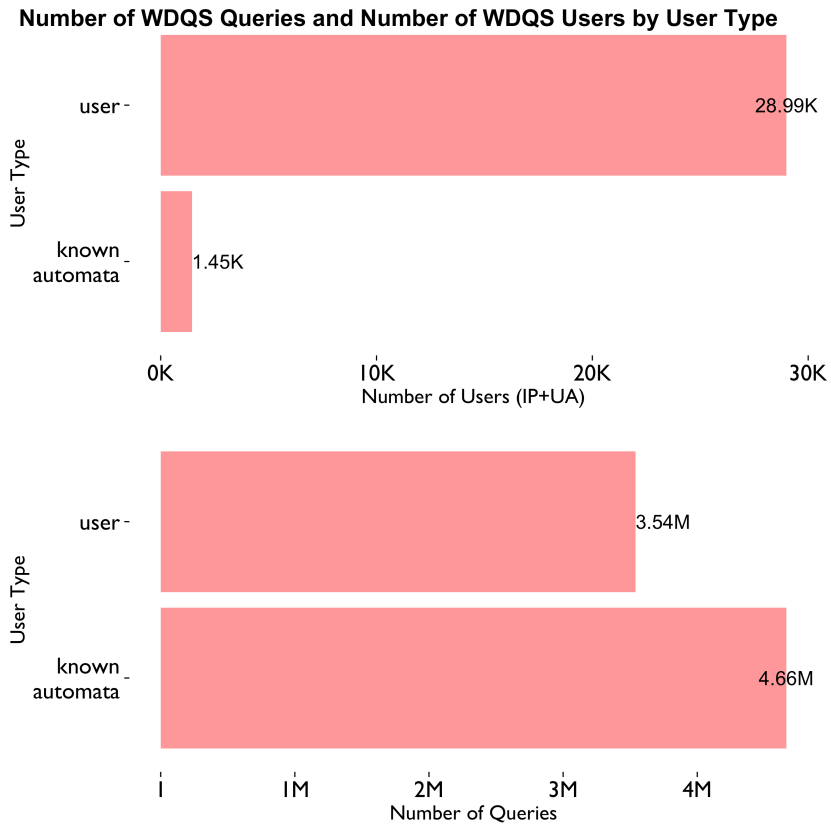

Figure 3: Number of WDQS Queries and Number of WDQS Users by User Type.

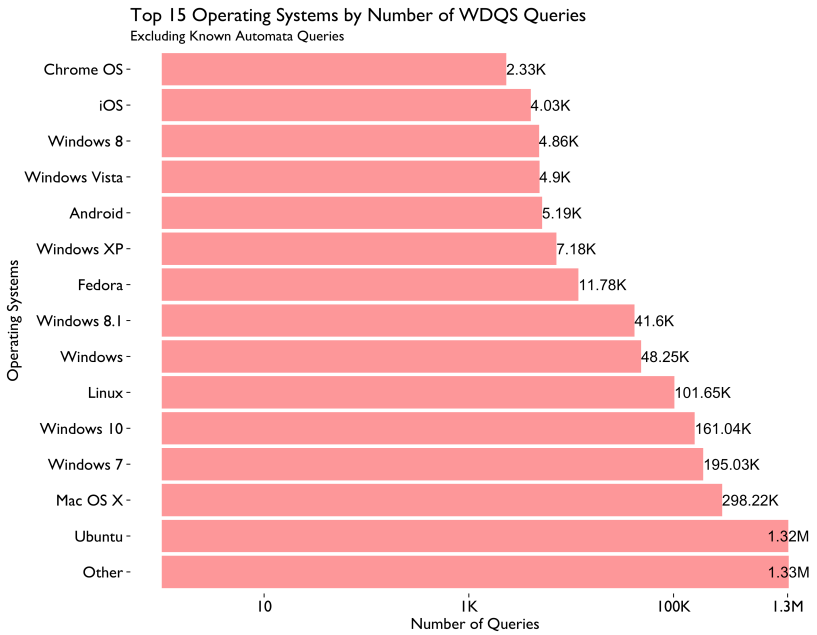

Figure 4: Top 15 Operating Systems by Number of WDQS Queries, Excluding Known Automata.

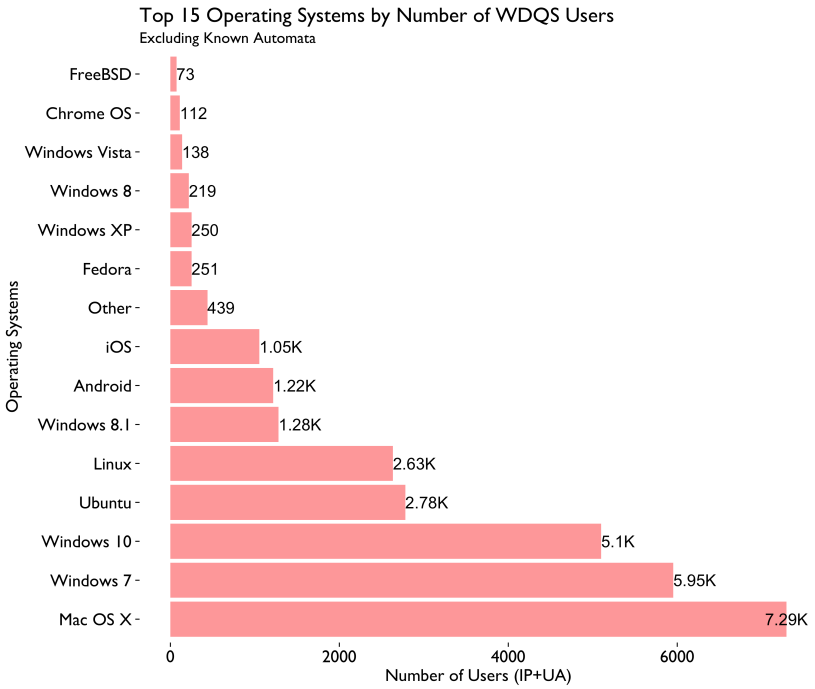

Figure 5: Top 15 Operating Systems by Number of WDQS Users, Excluding Known Automata. It looks like Ubuntu users have more queries-per-user than other operating systems on average.

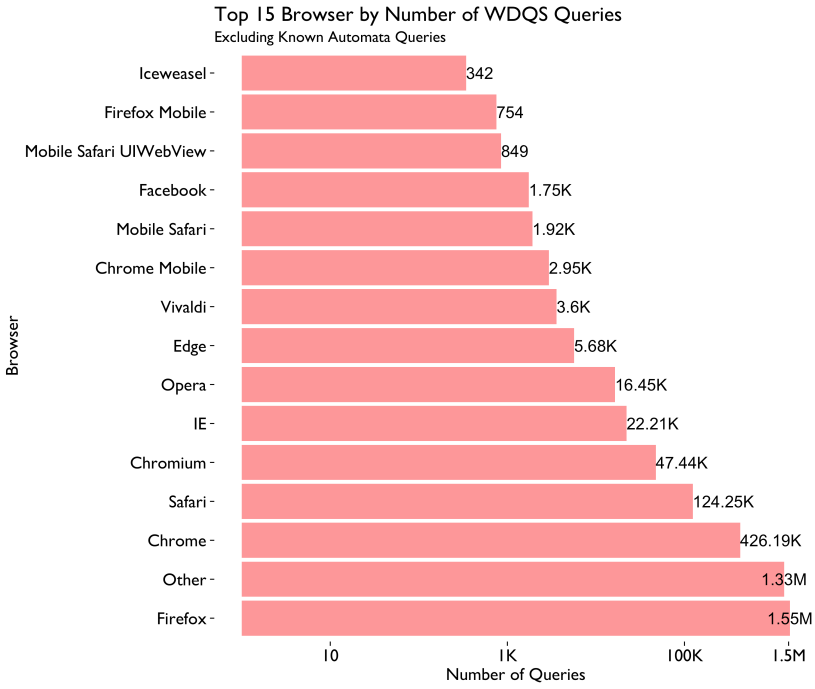

Figure 6: Top 15 Browser by Number of WDQS Queries, Excluding Known Automata.

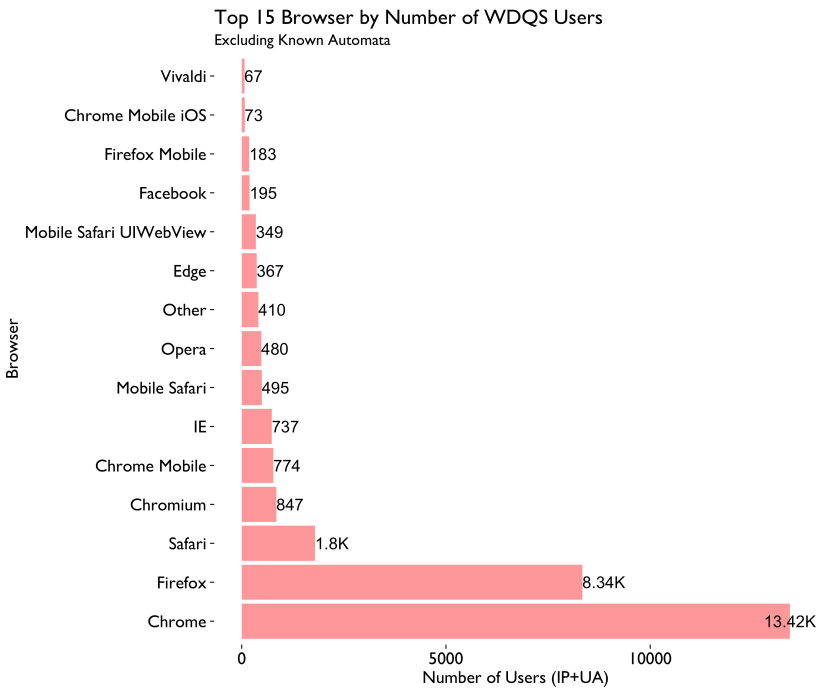

Figure 7: Top 15 Browser by Number of WDQS Users, Excluding Known Automata.

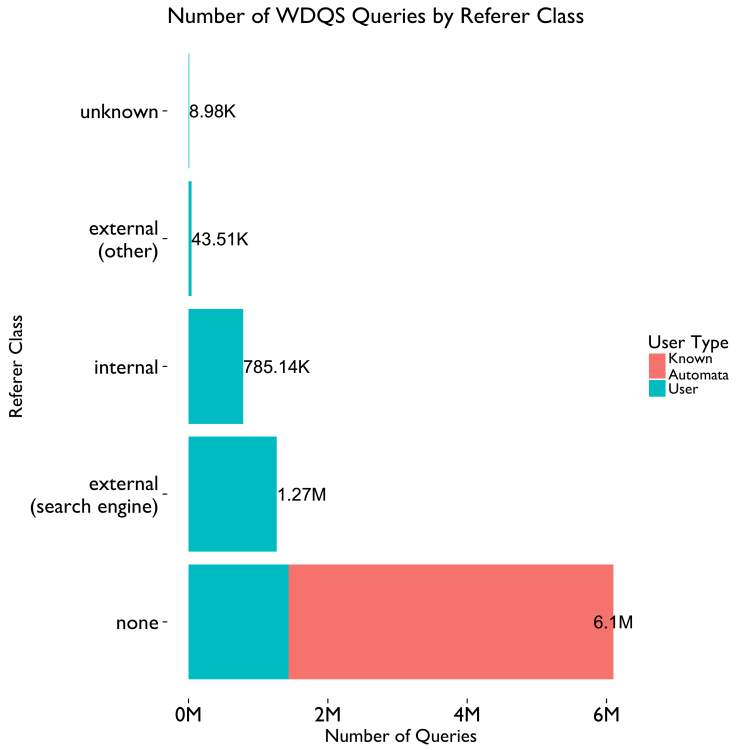

Figure 8: Most queries do not have a referer as they are from real people, but there are a large amount of queries that are from automata.

Longitudinal

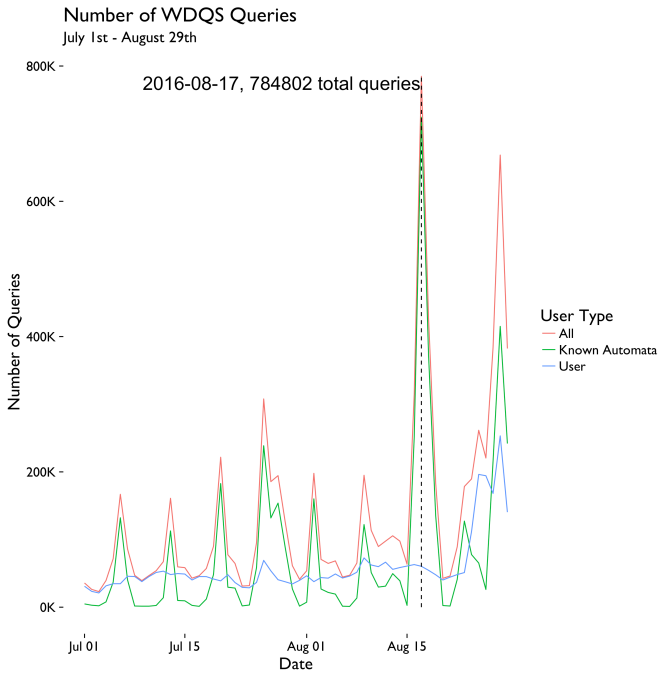

Figure 9: There seems to be a weekly cycle in the number of automata queries. After the spike (August 16-19), both types of queries saw an increase.

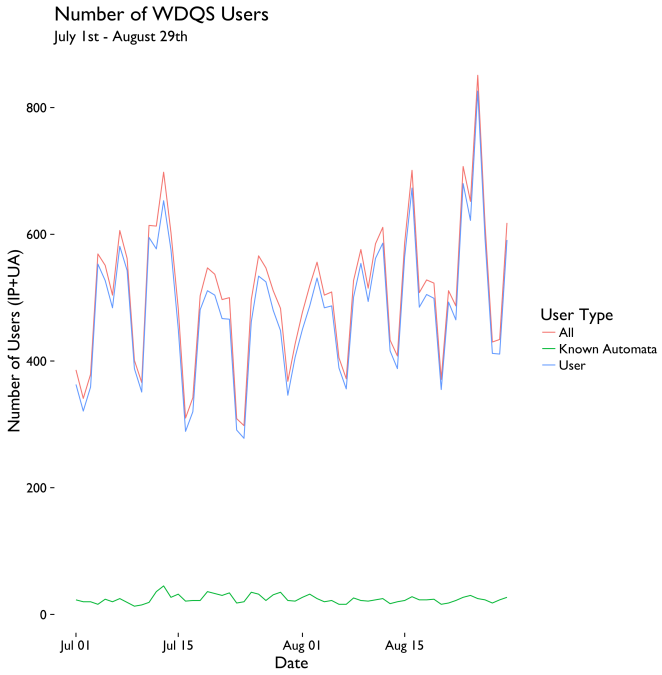

Figure 10: There seems to be a weekly cycle in the number of users.

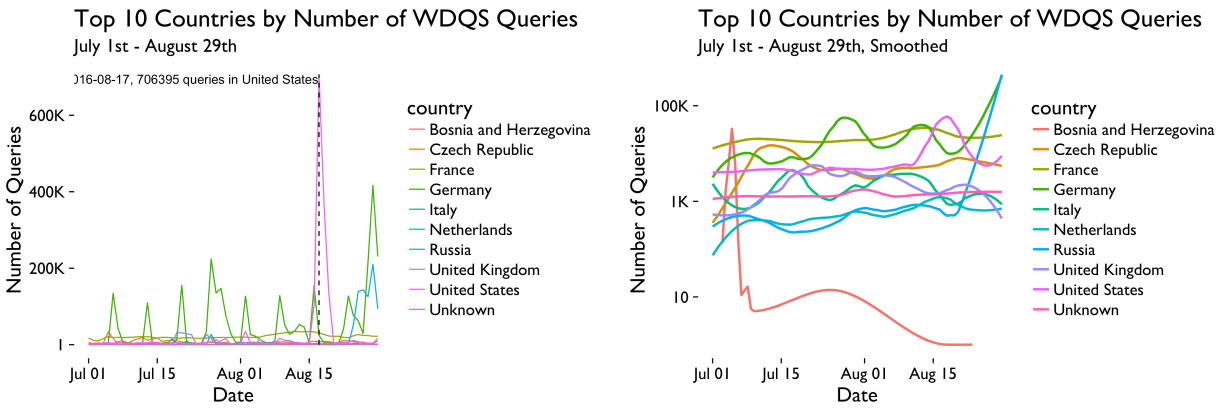

Figure 11: Further breakdown by country. The spike was contributed by the US. Germany seems to dominate the weekly cycle. Bosnia and Herzegovina saw a huge decrease in these two months.

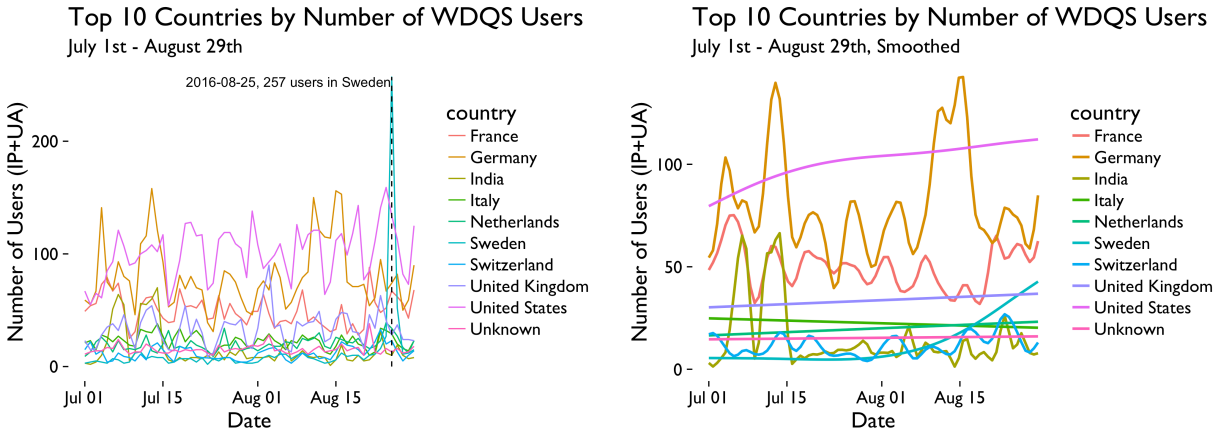

Figure 12: Top 10 Countries by Number of WDQS Users, July 1st - August 29th, 2016.

Next, we excluded the automata queries in US from August 16 to 19 (Figure 13), then implemented BFAST method on the query data. BFAST (Breaks For Additive Season and Trend) integrates the decomposition of time series into trend, season, and remainder components with methods for detecting and characterizing change within time series. First, it decomposes the series into trend and seasonal components with the STL method, then it uses OLS-MOSUM test on each components to see if there is any significant break point. Next, BFAST fits the two components and the detected break points with linear regression. BFAST iteratively estimates the time and number of changes, and characterizes change by its magnitude and direction, until the number and position of the breakpoints are unchanged.

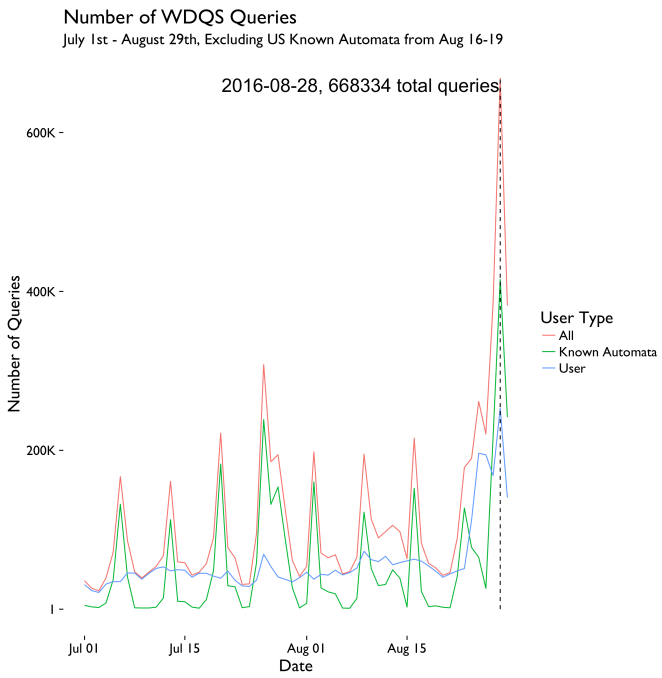

Figure 13: After excluding the automata queries from US Aug 16-19, the weekly cycle seems to hold for those days. Further investigation is needed to find out whether this spike is contributed by a particular automata.

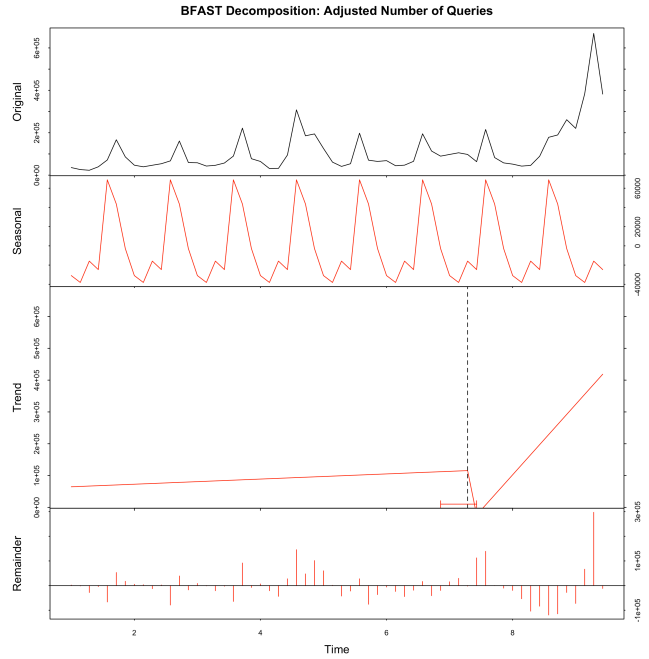

Figure 14: Adjusted number of queries decomposition. BFAST method detect a change point on Aug 14 in the trend component. At the change point, the decrease may be a result of our adjustment (excluding US automata), and more observations are needed to confirm the increase afterwards.

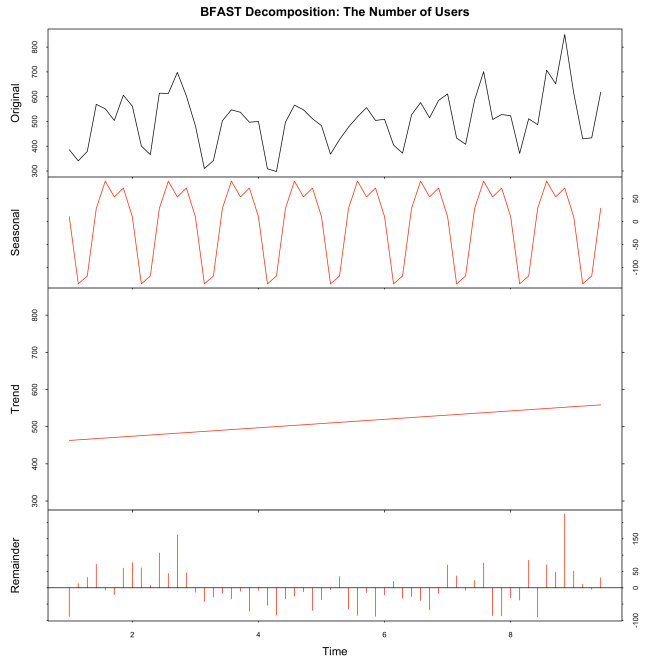

Figure 15: Number of users decomposition. There is no change point detected. We also see a slightly increasing trend.

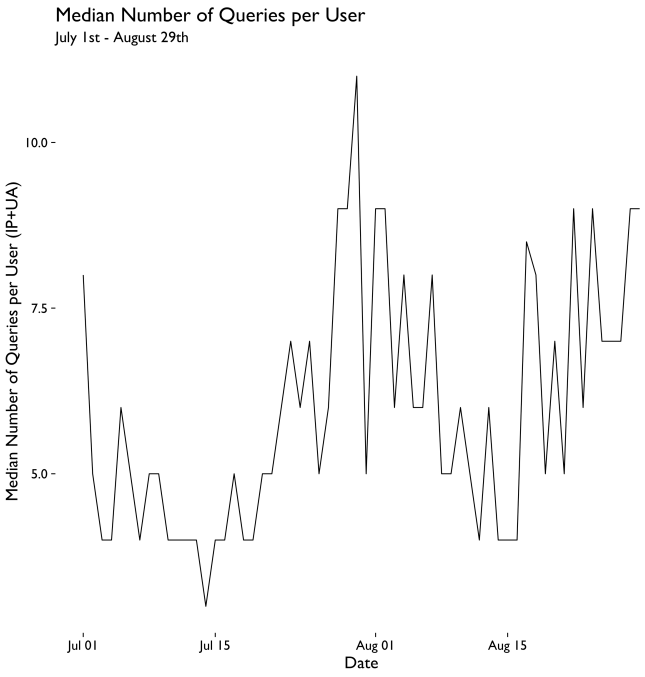

Figure 16: Median Number of Queries per User.

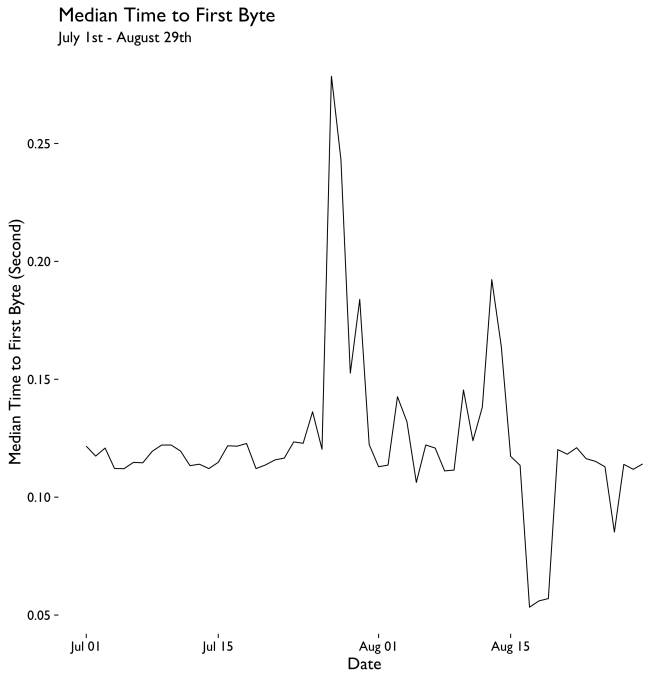

Figure 17: Time to first byte is a measurement used as an indication of the responsiveness of our servers. We saw a sharp decrease around August 17, which may be the result of the large number of automata queries. We also saw two spikes around July 25 and August 13, of which further investigation is needed.

Conclusion/Discussion

In summary, we found that:

- Germany, United States and France have the largest number of users and queries.
- Among regular users (not known automata), Mac OS X and Chrome are most popular, while Ubuntu and Firefox users submit the most queries.
- Most queries have no referer, followed by those referred from search engine.
- There are weekly cycles in the number of queries and users. And we also saw an increasing trend in the number of users.

For the next step, more thorough analysis and investigation are needed in order to solve the following questions:

- It looks like Ubuntu users have more queries-per-user than other operating systems on average. Is it the result of several outliers?
- We saw a large number of automata queries from US on Aug 16-19. Are they from one or several particular automata?
- What is the reason for the decrease number of queries of Bosnia and Herzegovina?
- Is there a significant increasing trend in the number of queries after August 19? Is the increasing trend in the number of users statistically significant?
- If we exclude known automata queries from Germany, could we still see the weekly cycle?
- What could possibly be the reason for the spikes and sharp decrease in the time to first byte pattern?