Where does Google find API documentation?

Treude, Christoph; Aniche, Mauricio

DOI
10.1145/3194793.3194796

Publication date
2018

Document Version
Publisher's PDF, also known as Version of record

Published in
2nd International Workshop on API Usage and Evolution

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.
Where does Google find API documentation?

Christoph Treude
University of Adelaide
Adelaide, Australia
christoph.treude@adelaide.edu.au

Mauricio Aniche
Delft University of Technology
Delft, the Netherlands
m.f.aniche@tudelft.nl

ABSTRACT
The documentation of popular APIs is spread across many formats, from vendor-curated reference documentation to Stack Overflow threads. For developers, it is often not obvious from where a particular piece of information can be retrieved. To understand this documentation landscape, we systematically conducted Google searches for the elements of ten popular APIs. We found that their documentation is widely dispersed among many sources, that GitHub and Stack Overflow play a prominent role among the search results, and that most sources are quick to document new API functionalities. These findings inform API vendors about where developers find documentation about their products, they inform developers about places to look for documentation, and they enable researchers to further study the software documentation landscape.

ACM Reference Format:

1 INTRODUCTION AND MOTIVATION
Many software development projects use libraries and frameworks whose functionality is made available through application programming interfaces (APIs) [13]. These APIs, such as the Java API, often come with curated documentation available on their websites. While this curated documentation can provide coherent and authoritative answers to many questions, the scope of such documentation is necessarily limited [8], and in many cases, the community has complemented this documentation with sources such as blogs [5], news aggregator discussions [1], and Stack Overflow threads [2].

For documentation consumers, it is often not obvious where a particular piece of information is stored [11]. Different documentation formats contain different kinds of information, written by different individuals and intended for different purposes [12]. For instance, the official documentation of an API typically captures information about functionality and structure, but lacks other types of information, such as concepts or purpose [3]. Some of the most severe obstacles faced by developers learning a new API are related to its documentation [7], in particular because of scarce information about the API’s design, rationale [6], usage scenarios, and code examples [7]. On the other hand, “how-to” questions [2] are the most frequent question type on Stack Overflow.

As a result of this dispersion of documentation, developers take to search engines to look for suitable documentation. To understand the resources that are available to developers when they search for API documentation on the Internet, in our earlier work from 2011 [4], we performed Google web searches for all API methods of one particular API—jQuery—and we examined the first ten search results for each API method. We found that 88% of the methods were covered by development blogs, mostly consisting of tutorials, and that 84% of the methods were covered on Stack Overflow.

The Internet is volatile: Web pages open and close, and the top search results returned for any given query change quickly. To keep up with these changes, in this paper, we present a replication of our work from 2011 for the jQuery API, and we complement this work with nine additional APIs. We also analyzed search results separately for API elements that had only been introduced recently. We find that in addition to the official documentation, search results from GitHub and Stack Overflow play a prominent role on the first page of results returned by Google. Interestingly, while search results from GitHub are more prominent than Stack Overflow for some APIs (e.g., Tensorflow), the opposite is true for other APIs (e.g., jQuery). For some APIs (e.g., Hadoop), the API’s issue tracker is featured prominently among the search results, while for others (e.g., Guava, JUnit), a tutorial site with paid content is frequently returned by Google. As an example of the changes since 2011, GitHub—which we only mentioned as a side note in our earlier work—is now among the top five domains for all ten APIs that we considered in this study.

2 METHODOLOGY
We ask two research questions:

RQ1. Where does Google find API documentation?

RQ2. Do resources found for recently introduced API elements differ from the rest?

Answers to the first research question will help characterize the documentation landscape and its dispersion for different APIs, while answers to the second research question will inform developers about which documentation sources might be slow to document new API functionalities.

Data Collection. To answer our research questions, we selected ten popular APIs, aiming to cover a variety of programming languages and sizes. Table 1 lists the selected APIs along with their programming language and the API versions used in this work. The APIs span five programming languages. For each API, we determined when the most recent API version had been released at the
We then queried Google through a Google Custom Search Engine (https://cse.google.com/cse/all) and the Google Custom Search JSON API (https://developers.google.com/custom-search/json-api/v1/using_rest) to answer our second research question about where resources found for recently introduced API elements differ from the rest, we repeated the analysis described in the previous paragraph, but only for API elements that were available in the most recent API version but not in the previous one, as per the version numbers in Table 1. For our second research question, i.e., do resources found for recently introduced API elements differ from the rest, we repeated the analysis described in the previous paragraph, but only for API elements that were available in the most recent API version but not in the previous one, as per the version numbers in Table 1. All raw and aggregated data are available online.\footnote{See our online appendix for details on a small subset that we excluded.}

\section{3 FINDINGS}

\subsection{Sources of API documentation}

Table 2 shows the total number of domains from which search results originated, separately for each API (domains). The numbers demonstrate that API documentation is widely dispersed among many domains, e.g., the 5,693 searches for the Java API returned results from 4,139 domains on the first page of search results alone. While there is a strong correlation (Pearson’s \( r = 0.94 \)) between the size of an API measured in terms of its number of elements (and consequently the number of queries we conducted) and the number of domains, the documentation of some APIs is more dispersed than that of other APIs: Documentation for the 226 classes of JUnit can be found on 252 domains when only considering the first page of Google search results—in other words, there are more domains than API elements in this case. We define the \textit{documentation dispersion factor} of an API as the number of the search results returned by Google. Note that in some cases, the number of links returned is not exactly ten—it might be higher if Google identified multiple links belonging to a single site and displayed them as sub-links to one higher-level search result, or it might be lower if Google found fewer than ten results in total for a particular query. For each link, we also identified its rank in the list of search results.

\textbf{Data Analysis.} To answer our first research question, i.e., where does Google find API documentation, we determined the domain of each link retrieved in the previous step, and for each domain, we determined its coverage and median rank with regard to a specific API. We define \textit{coverage} as the percentage of API elements for which a particular domain appeared on the first page of Google search results, and we define \textit{median rank} as the median of all ranks of a particular domain when it appeared on the first page of the Google search results. Note that if a domain appeared more than once on the first page of the Google search results for a single query, we only considered the link with the highest rank for the calculation of the median rank across all queries.

For our second research question, i.e., do resources found for recently introduced API elements differ from the rest, we repeated the analysis described in the previous paragraph, but only for API elements that were available in the most recent API version but not in the previous one, as per the version numbers in Table 1.

All raw and aggregated data are available online.\footnote{http://doi.org/10.5281/zenodo.1195863}

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|}
\hline
API & language & current & previous & url \\
\hline
Guava & Java & 230 (4-Aug-17) & 22.0 (22-May-17) & https://google.github.io/guava/releases/23.0/api/docs/allclasses-noframe.html \\
Hadoop & Java & 3.0.0 (13-Dec-17) & 2.7.4 (4-Aug-17) & https://hadoop.apache.org/docs/r3.0.0/api/docs/allclasses-noframe.html \\
Java & Java & 9 (21-Sep-17) & 8 (18-Mar-14) & https://docs.oracle.com/javase/9/api/docs/allclasses-noframe.html \\
jQuery & JavaScript & 3.2.1 (21-Mar-17) & 3.0 (10-Jun-16) & https://api.jquery.com/ \\
JUnit & Java & 5.0.2 (12-Nov-17) & 4.12 (5-Dec-14) & http://junit.org/junit5/docs/current/api/allclasses-noframe.html \\
Laravel & PHP & 5.5 (30-Aug-17) & 5.4 (24-Jan-17) & https://laravel.com/api/5.5/classes.html \\
Qt & C++ & 5.10 (7-Dec-17) & 5.8 (23-Jan-17) & http://doc.qt.io/qt-5/classes.html \\
Symfony & PHP & 4.0.1 (30-Nov-17) & 3.4 (29-May-17) & https://api.symfony.com/4.0/classes.html \\
Tensorflow & Python & 1.4 (3-Nov-17) & 1.3 (17-Aug-17) & https://www.tensorflow.org/api_docs/python/ \\
\hline
\end{tabular}
\caption{APIs used in our study}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|c|}
\hline
API & total & recent & & & \\
& elems. & domains & dom. & elems. & domains \\
\hline
JUnit & 226 & 252 & 1.12 & 224 & 249 \\
jQuery & 296 & 249 & 0.84 & 3 & 12 \\
Guava & 399 & 320 & 0.80 & 3 & 8 \\
Android & 4,140 & 3,196 & 0.77 & 18 & 57 \\
Java & 5,693 & 4,139 & 0.73 & 1,589 & 1,947 \\
Hadoop & 826 & 594 & 0.72 & 172 & 181 \\
Laravel & 675 & 486 & 0.72 & 53 & 93 \\
Symfony & 1,700 & 738 & 0.43 & 113 & 83 \\
Qt & 1,609 & 524 & 0.33 & 36 & 28 \\
Tensorflow & 2,582 & 583 & 0.23 & 826 & 253 \\
\hline
\end{tabular}
\caption{Table 2: Documentation dispersion}
\end{table}

We then queried Google through a Google Custom Search Engine\footnote{https://cse.google.com/cse/all} and the Google Custom Search JSON API\footnote{https://developers.google.com/custom-search/json-api/v1/using_rest} with each API element separately, prefixing each query with the name of the corresponding API (e.g., we searched for “Java ArrayList” and “jQuery .add()”). The Google Custom Search Engine was configured to search the entire web, and we did not specify any particular sites to be included. We then retrieved all links from the first page of the search results returned by Google. Note that in some cases, the number of links returned is not exactly ten—it might be higher if Google identified multiple links belonging to a single site and displayed them as sub-links to one higher-level search result, or it might be lower if Google found fewer than ten results in total for a particular query. For each link, we also identified its rank in the list of search results.
Table 3: Top domains for documentation (rk. = median rank) of domains divided by the number of elements, shown in Table 2 (domains). While many APIs have a factor in the range between 0.72 and 0.84, JUnit is an outlier with a high factor and TensorFlow, Qt, and Symfony are outliers with a low factor, suggesting that these APIs are documented on a relatively small set of domains. Note that even these APIs still resulted in at least 500 domains.

Table 4: Top domains for documentation (rk. = median rank) documentation achieved the highest coverage with values above 97% except Symfony (92.9%). We speculate that the ambiguity of the name of the API explains the lower coverage. For all APIs, search results from GitHub and Stack Overflow played a prominent role on the first page of search results returned by Google. Whether GitHub or Stack Overflow is a more important resource for API documentation depends on the API: search results from GitHub were more prominent than Stack Overflow for some APIs (e.g., TensorFlow), while the opposite was true for other APIs (e.g., jQuery).

Other domains that entered the top five include Google’s Git repository hosting site android.googlesource.com and the Xamarin developer center developer.xamarin.com for Android, the web development tutorial site with paid content baeldung.com, and the Javadoc for scientific computing hosting site javadoc.scijava.org for Guava, as well as the Hadoop issue.
When we compared each domain in terms of its coverage of all API elements and its coverage of recent API elements, we found that documentation is dispersed among many sources with GitHub and Stack Overflow playing prominent roles, and that most sources are quick to document new API functionalities. These findings support API vendors and users by characterizing the documentation landscape, and the data available in our appendix enables researchers to further study the dispersion of API documentation.

4 CONCLUSIONS

To understand which resources developers find when searching for API documentation, we systematically performed web searches for the elements of ten popular APIs. We found that documentation is dispersed among many sources with GitHub and Stack Overflow playing prominent roles, and that most sources are quick to document new API functionalities. These findings support API vendors and users by characterizing the documentation landscape, and the data available in our appendix enables researchers to further study the dispersion of API documentation.

REFERENCES