Where does Google find API documentation?

Treude, Christoph; Aniche, Maurício

DOI
10.1145/3194793.3194796

Publication date
2018

Document Version
Final published version

Published in
2nd International Workshop on API Usage and Evolution

Where does Google find API documentation?

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

Maurício 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:
Christoph Treude and Maurício Aniche. 2018. Where does Google find API documentation?. In *WAPI'18: WAPI'18: IEEE/ACM 2nd International Workshop on API Usage and Evolution, June 2–4, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3194793.3194796

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

WAPI'18, June 2–4, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5754-8/18/06…$15.00
https://doi.org/10.1145/3194793.3194796

API	language	current	previous	url
Android	Java	27 (5-Dec-17)	26 (21-Aug-17)	https://developer.android.com/reference/classes.html
Guava	Java	23.0 (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/allclasses-noframe.html
Java	Java	9 (21-Sep-17)	8 (18-Mar-14)	https://docs.oracle.com/javase/9/docs/api/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/

Table 1: APIs used in our study

	total			recent	
API	elems.	domains	$\frac{\text{domains}}{\text{element}}$	elems.	domains
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

Table 2: Documentation dispersion

time of data collection. Table 1 shows the corresponding version number (*current*) and the URL from which the elements of each API were retrieved (*url*). To answer our second research question about recently introduced API elements, we retrieved an earlier version for each API. We tried to identify API versions that were only a few months older than the current version. As Table 1 shows, this was not possible in all cases since some of the selected APIs do not get updated frequently.

For each API, we then identified its constituents, i.e., the elements that make up the API. For Java-based APIs, these were the classes provided by each API.[1] For example, version 9 of the Java API exposes 5,693 classes while JUnit 5.0.2 provides 226 classes. For jQuery, we used the JavaScript methods made available through its API as API elements.[2] For the Tensorflow Python API, we used its 2,582 symbols as API elements. Finally, for the APIs written for PHP (Laravel and Symfony) and the API written for C++ (Qt), we used classes as API elements, similar to the Java APIs. Table 2 contains the number of API elements we identified for each API (*elems.*).

We then queried Google through a Google Custom Search Engine[3] and the Google Custom Search JSON API[4] 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.

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 *coverage* as the percentage of API elements for which a particular domain appeared on the first page of Google search results, and we define *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.[5]

3 FINDINGS

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 *documentation dispersion factor* of an API as the number

[1] Conducting an analysis on API methods of Java-based APIs is part of our future work.
[2] See our online appendix for details on a small subset that we excluded.
[3] https://cse.google.com/cse/all
[4] https://developers.google.com/custom-search/json-api/v1/using_rest
[5] http://doi.org/10.5281/zenodo.1195863

Android	total		recent	
(4,140 total, 18 recent)	%	rk.	%	rk.
developer.android.com	99.5%	1	94.4%	1
stackoverflow.com	85.1%	2	61.1%	4
github.com	59.8%	6	61.1%	6
android.googlesource.com	44.4%	5	38.9%	2
developer.xamarin.com	40.3%	6	–	–
Guava	total		recent	
(399 total, 3 recent)	%	rk.	%	rk.
google.github.io	100.0%	1	100.0%	1
github.com	96.5%	2	100.0%	2
stackoverflow.com	88.5%	5	33.3%	3
baeldung.com	41.6%	5	–	–
javadoc.scijava.org	37.6%	7	–	–
Hadoop	total		recent	
(826 total, 172 recent)	%	rk.	%	rk.
hadoop.apache.org	99.2%	1	98.8%	1
stackoverflow.com	52.3%	4	31.4%	5
issues.apache.org	47.7%	4	41.3%	2
archive.cloudera.com	34.5%	5	7.6%	6
github.com	33.9%	5	20.3%	6
Java	total		recent	
(5,693 total, 1,589 recent)	%	rk.	%	rk.
docs.oracle.com	97.7%	1	93.0%	1
stackoverflow.com	77.1%	3	74.0%	3
github.com	38.0%	6	41.0%	6
java2s.com	22.6%	5	14.3%	6
ibm.com	12.0%	6	7.5%	6
jQuery	total		recent	
(296 total, 3 recent)	%	rk.	%	rk.
api.jquery.com	100.0%	1	100.0%	1
stackoverflow.com	89.9%	4	100.0%	5
w3schools.com	79.7%	2	–	–
github.com	45.6%	7	100.0%	3
learn.jquery.com	18.2%	6	–	–

Table 3: Top domains for documentation (*rk.* = median rank)

JUnit	total		recent	
(226 total, 224 recent)	%	rk.	%	rk.
junit.org	98.7%	1	98.7%	1
github.com	80.5%	2	80.8%	2
stackoverflow.com	65.9%	4	66.1%	4
blog.codefx.org	17.7%	6	17.9%	6
baeldung.com	15.9%	4.5	16.1%	4.5
Laravel	total		recent	
(675 total, 53 recent)	%	rk.	%	rk.
laravel.com	97.2%	1	83.0%	1
github.com	88.1%	4	66.0%	2
stackoverflow.com	78.1%	4	62.3%	4
laracasts.com	75.1%	5	58.5%	4
laravel-news.com	16.7%	5	30.2%	3
Qt	total		recent	
(1,609 total, 36 recent)	%	rk.	%	rk.
doc.qt.io	100.0%	1	91.7%	1
stackoverflow.com	69.2%	4	5.6%	7
archlinux.org	32.5%	5	69.4%	2
github.com	32.0%	6	2.8%	6
pyqt.sourceforge.net	27.3%	6	5.6%	6
Symfony	total		recent	
(1,700 total, 113 recent)	%	rk.	%	rk.
api.symfony.com	92.9%	2	95.6%	2
github.com	89.8%	2	70.8%	1
stackoverflow.com	73.5%	4	35.4%	4
symfony.com	73.4%	2	46.0%	1
knpuniversity.com	15.1%	6.5	1.8%	4.5
Tensorflow	total		recent	
(2,582 total, 826 recent)	%	rk.	%	rk.
tensorflow.org	99.7%	1	99.3%	1
github.com	88.6%	2	82.2%	4
stackoverflow.com	69.6%	4	55.3%	6
w3cschool.cn	24.5%	6	5.2%	9
keras.io	17.6%	2	53.1%	2

Table 4: Top domains for documentation (*rk.* = median rank)

of domains divided by the number of elements, shown in Table 2 ($\frac{domains}{element}$). 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.

Tables 3 and 4 show the top domains for each API along with the domains' coverage and the median rank. For example, at least one search result from the domain `developer.android.com` appeared on the first page of Google search results for 99.5% of the 4,140 queries related to the Android API, and the median rank of the first search result from this domain was 1. The domain `stackoverflow.com` was ranked second in terms of coverage (85.1%) at a median rank of 2, and `github.com` came in third with a coverage of 59.8% and a median rank of 6. For all APIs, their official 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

API	domain	total	recent	diff
Qt	archlinux.org	32.5%	69.4%	+36.9%
Tensorflow	keras.io	17.6%	53.1%	+35.5%
Hadoop	archive.cloudera.com	34.5%	7.6%	-26.9%
Symfony	symfony.com	73.4%	46.0%	-27.4%
Qt	github.com	32.0%	2.8%	-29.2%
Symfony	stackoverflow.com	73.5%	35.4%	-38.1%
Android	developer.xamarin.com	40.3%	0.0%	-40.3%
Qt	stackoverflow.com	69.2%	5.6%	-63.6%

Table 5: Differences in coverage for recently added elements

tracker at `issues.apache.org` and the archive of the Cloudera education site at `archive.cloudera.com` for Hadoop. For Java, the top five includes the programming tutorial and source code example site `java2s.com` along with `ibm.com`, and for jQuery, we found the learning, testing, and training site for web developers `w3schools.com` in the top five along with the jQuery learning center at `learn.jquery.com`. The code blog `blog.codefx.org` and `baeldung.com` are featured prominently for JUnit, while Laravel documentation can be found on the news site `laravel-news.com` and in the form of screencasts on `laracasts.com`. For Qt, the domain of Arch Linux, a lightweight Linux distribution, at `archlinux.org` is commonly found on the first page of Google search results, along with the domain for Python bindings for the Qt application framework at `pyqt.sourceforge.net`. For Symfony, the tutorial site `knpuniversity.com` is prominent among the search results, while the Chinese tutorial site `w3cschool.cn` and the neural networks API Keras at `keras.io` complete the top five for Tensorflow. We refer readers to our online appendix for a complete list of domains along with coverage and median ranks.

GitHub accounts for the largest difference in coverage when comparing the results from this study to the original study on jQuery documentation [4]: it was only mentioned as a side note in 2011 and now covers 45.6% of the jQuery API methods. The official API documentation has remained the most prominent source of search results, while Stack Overflow (from 84.4% to 89.9%) and unofficial documentation sources such as `w3schools.com` (from 63.6% to 79.7%) have risen slightly in terms of their coverage. On the other hand, blog posts and the official jQuery forums appear to play a less important role now.

Documentation of recently added elements. When we compared each domain in terms of its coverage of all API elements and its coverage of recent API elements, we did not find many differences. This finding suggests that most sources which cover API documentation are quick to document new API functionalities, and that Google is quick to include these additions in its results.

Table 5 shows the only eight domains in our dataset for which the difference in coverage between all API elements and recent API elements exceeded 25%. Note that we excluded jQuery and Guava from this analysis since these APIs only had a small number of recent elements (cf. Table 2). Two domains, `archlinux.org` and `keras.io`, achieved a much higher coverage for recently added API elements compared to the rest, suggesting that these domains are particularly fast to document new functionality. Six domains achieved a much lower coverage. Particularly noteworthy is the case of the Qt API, for which the coverage of the 36 API classes added in January 2017 on GitHub and Stack Overflow was less than 6%—these two domains had a coverage of 32.0% and 69.2%, respectively, when considering all 1,609 classes of the Qt API. Another noteworthy finding affects the Android API: None of the 18 classes added in August 2017 resulted in search results from `developer.xamarin.com`, compared to an overall coverage of 40.3%. Analyzing the documentation of deprecated API elements [9] is part of our future work.

Threats to Validity. Since Google does not expose search results through an API, we had to rely on a Google Custom Search Engine. We manually verified that the results from `google.com` and from our Google Custom Search Engine were almost identical, but we cannot guarantee that a similar analysis conducted manually on `google.com` would find the exact same results. While considering all results beyond the first page of search results would affect the coverage values reported here, users rarely investigate results that are not on the first page [10]. We only conducted queries which combined the name of an API with exactly one API element—this might not be an accurate representation of typical developer queries.

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

[1] M. Aniche, C. Treude, I. Steinmacher, I. Wiese, G. H. L. Pinto, M.-A. Storey, and M. A. Gerosa. 2018. How Modern News Aggregators Help Development Communities Shape and Share Knowledge. In *Int'l. Conf. on Software Engineering*.

[2] O. Barzilay, C. Treude, and A. Zagalsky. 2013. Facilitating Crowd Sourced Software Engineering via Stack Overflow. In *Finding Source Code on the Web for Remix and Reuse*, S. E. Sim and R. E. Gallardo-Valencia (Eds.). Springer, New York, United States, 289–308.

[3] W. Maalej and M. P. Robillard. 2013. Patterns of Knowledge in API Reference Documentation. *IEEE Trans. on Software Engineering* 39, 9 (2013), 1264–1282.

[4] C. Parnin and C. Treude. 2011. Measuring API Documentation on the Web. In *Int'l. Workshop on Web 2.0 for Software Engineering*. 25–30.

[5] C. Parnin, C. Treude, and M.-A. Storey. 2013. Blogging developer knowledge: Motivations, challenges, and future directions. In *Int'l. Conf. on Program Comprehension*. 211–214.

[6] M. P. Robillard. 2009. What Makes APIs Hard to Learn? Answers from Developers. *IEEE Software* 26, 6 (2009), 27–34.

[7] M. P. Robillard and R. DeLine. 2011. A Field Study of API Learning Obstacles. *Empirical Software Engineering* 16, 6 (2011), 703–732.

[8] M. P. Robillard, A. Marcus, C. Treude, G. Bavota, O. Chaparro, N. Ernst, M. A. Gerosa, M. Godfrey, M. Lanza, M. Linares-Vásquez, G. Murphy, L. Moreno, D. Shepherd, and E. Wong. 2017. On-Demand Developer Documentation. In *Int'l. Conf. on Software Maintenance and Evolution*. 479–483.

[9] A. Sawant, M. Aniche, A. van Deursen, and A. Bacchelli. 2018. Understanding Developers' Needs on Deprecation as a Language Feature. In *Int'l. Conf. on Software Engineering*.

[10] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (1999), 6–12.

[11] C. Treude and M. P. Robillard. 2016. Augmenting API Documentation with Insights from Stack Overflow. In *Int'l. Conf. on Software Engineering*. 392–403.

[12] C. Treude and M.-A. Storey. 2011. Effective Communication of Software Development Knowledge Through Community Portals. In *Symp. and the European Conf. on Foundations of Software Engineering*. 91–101.

[13] G. Uddin and M. P. Robillard. 2015. How API Documentation Fails. *IEEE Software* 32, 4 (2015), 68–75.