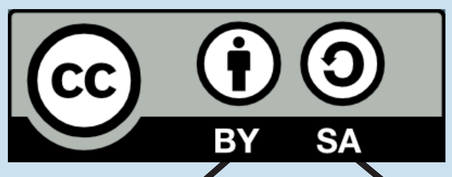
Usage and Attribution of

Code Snippets in GitHub Projects

Universität Trier
Software Engineering Group

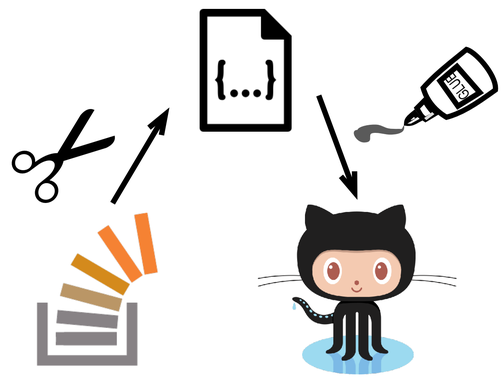
License of Stack Overflow Content



Attribution: Developers using code snippets from Stack Overflow (SO) must attribute author and origin of the snippet.

Share Alike: Derived work must be distributed under a compatible license.

Usage without attribution leads to legal and maintenance issues.



Research Questions

RQ1 How is content from Stack Overflow referenced in GitHub projects?

Method: Searched source code of 20.2m GitHub (GH) repos for SO URLs using BigQuery.

Results:
- About twice as many references to questions than to answers.
- On average, 3.22% of all repos and 7.33% of the popular ones contained a reference to SO.
- R, Python, C#, and Objectice-C files contained considerably more references to SO than the other analyzed languages.

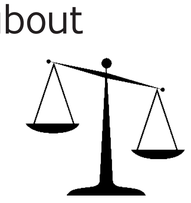
RQ2 How often is code from Stack Overflow posts used, but not attributed?

Method: Searched for duplicates of SO snippets using a code clone detector and regular expressions.

Results:
- Searching for code clones of SO snippets in a sample of GH Java projects (n=2,313) revealed that only 23% of the clones were attributed.
- Using BigQuery and regexes, we searched for duplicates of snippets from ten most frequently referenced SO Java answers in 336,028 GH projects; only 27% of the identified usages were attributed.

RQ3 What distinguishes frequently from less frequently referenced Stack Overflow content?

Method: Retrieve and analyze information about references extracted for RQ1 using SO API.



Results:
- Frequently referenced questions and answers had a significantly higher view count and score,
- Frequently referenced answers had significantly more code blocks, but the effect was only small.

RQ4 Are software developers aware of the licensing situation of Stack Overflow code snippets?

Method: Online survey with 87 developers having duplicates of SO code snippets in their GH repo(s).

Results:
- Most developers (75%) were not aware of the licensing of code published on SO.
- Only 32% of the developers were aware of the attribution requirement for content from SO.

Sebastian Baltes
research@sbaltes.com
@s_baltes

Stephan Diehl
diehl@uni-trier.de

http://snippets.sbaltes.com

39th International Conference on Software Engineering
May 20-28, 2017 - Buenos Aires, Argentina