# GitHub Actions Developer Information Needs: An Empirical Study

1st Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

2nd Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

3rd Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

4th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

5th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

6th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.** TO DO ►*Lo escribimos al final*◄

***Index Terms*—component, formatting, style, styling, insert**

## I. Introduction

The software industry has widely adopted Continuous Integration and Delivery (CI/CD) practices to automate software engineering tasks [1]. These practices help minimize integration issues, enable frequent integration, automatically deploy changes, and speed up feedback loops for software developers [2]–[4].

Within the GitHub ecosystem, developers can automate software engineering tasks through GitHub Actions (GA), a widely adopted tool for implementing CI/CD pipelines [1]. GA provides technical mechanisms to define, execute, and manage workflows using `YAML` files, allowing developers to specify automated tasks triggered by events such as code pushes, and pull requests.

While GA provides multiple mechanisms to automate software engineering tasks, developers still need specific information when using this tool, as evidenced by their technical questions on discussion forums like Stack Overflow (SO). For instance, the SO post titled "How to implement semantic versioning in GitHub Actions workflow"[1] reflects the author's search for a way to implement semantic versioning in a GitHub Actions workflow. Semantic versioning is the version numbering system that follows the `MAJOR.MINOR.PATCH` format[2]. The author explains that he has a GitHub repository and needs to automate versioning following this convention using GA, as the tool allows this directly on GitHub. Despite finding some examples and documentation, the author has not been able to find a complete solution that meets their needs. This example not only reflects the need for specific information but also aligns with previous studies, which show that developers often seek to understand how to implement specific tasks, indicating a need for additional information that is not readily available and prevents them from moving forward [5]. These observations suggest that the information provided by current tools does not match the knowledge developers require to make informed decisions.

These observations suggest that the information provided by the GA tool does not match the knowledge developers need to make informed decisions. Although research studies have extensively explored developer information needs [6], [7], even in the context of similar software artifacts [8], [9], they have not been conducted specifically related to GA.

We present an empirical study aimed at characterizing the GA developer information needs. We define GA developer information needs as the questions asked on SO by individuals from various backgrounds (e.g., students, professional developers, etc.), whom we refer to simply as developers. Our study is based on XXX SO posts associated with GA, involving XXX specific sentences that highlight developer information needs. The study focuses on XXX.

PV ►*still working on this section*◄

TO DO ►*Sandro, completar esta sección*◄

## II. Study Design

The goal of this study is to characterize the GA developer information needs. The purpose is to define a taxonomy of information needs that developers expose in the SO platform. The study address the following Research Questions (RQ):

- **RQ1: What is the current level of interest in the topic?** This question aims to identify the trend in the number of SO posts related to GitHub Actions and compare it with the trends of other popular tools in Continuous

[1] https://stackoverflow.com/questions/70925425/how-to-implement-semantic-versioning-in-github-actions-workflow

[2] https://semver.org/

Integration. Through this comparison, we will determine the level of interest in GitHub Actions.

- **RQ2: What types of information needs are presented in SO posts about GitHub Actions?** This research question seeks to characterize the types of information needs that developers have. By categorizing the questions and discussions, we can better understand the common information needs developers face when using GitHub Actions.

TO DO ►*Sandro, usando la estructura de abajo, por favor completar detalladamente desde Data Collection - Analysis Techniques. No incluir los resultados acá. Este paper paper es un ejmeplo del detalle que necesitamos acá: https://rebels.cs.uwaterloo.ca/journalpaper/2023/09/10/an-empirical-comparison-of-ethnic-and-gender-diversity-of-devops-and-non-devops-contributions-to-open-source-projects.html*◄

## A. RQ1: Level of interest

### 1) Data Curation

To address the level of interest in GitHub Actions, we first identified the most commonly used tools for Continuous Integration (CI). According to the results of a survey conducted by JetBrains TeamCity in 2023, the six most popular CI tools are Jenkins, GitHub Actions, GitLab CI, Azure DevOps, CircleCI, and Travis CI.

We identified the most significant tags related to each of these tools on StackOverflow, which are: ‘github-actions’, ‘jenkins’, ‘gitlab-ci’, ‘azure-devops’, ‘circleci’, and ‘travis-ci’. Although there are additional derivative tags related to these tools, our analysis focuses on the primary tags to effectively compare the trends in the posts. We extracted the monthly number of posts tagged with the most popular CI tools from the beginning of 2019 to October 2023 to be compared.

This dataset includes a total of 60,389 posts, distributed as follows: 9622 with the tag ‘github-actions’; 19354 with ‘jenkins’; 6581 with ‘gitlab-ci’; 22797 with ‘azure-devops’; 1095 with ‘circleci’; and 940 with ‘travis-ci’.

## B. RQ2: Types of Information Needs

### 1) Data Curation

Our data collection concentrated on retrieving SO posts related to GitHub Actions, with a particular focus on filtering posts that explicitly discuss this tool. We utilized the StackOverflow Data Dump to extract these posts and their metadata, covering entries up until the end of 2023. To ensure the relevance of the collected posts, we applied several filtering techniques. Following this, we selected a representative sample of the posts and systematically coded the text. Below, we detail each step of this process.

1) **Tag Filtering:** Posts were filtered using tags associated with GitHub Actions. These tags were identified through a search for ”github actions” in the StackOverflow tag search bar, resulting in a list that included ’github-actions’, ’building-github-actions’, ’github-actions-self-hosted-runners’, ’github-actions-runners’, ’github-actions-services’, ’github-actions-artifacts’, ’github-actions-reusable-workflows’, ’github-actions-workflows’, and ’github-actions-marketplace’.
2) **Title and Body Filtering:** We also scrutinized the post titles or bodies for mentions of GitHub Actions or its variations. The rules that we established for variations of a word or phrase were considering various capitalizations of the first letter of the phrase or word, and hyphenations (’github actions’, ’github-actions’, ’Github actions’, ’Github-actions’, ’Github Actions’, ’Github-Actions’).
3) **Manual Filtering:** A manual inspection was applied to ensure that the post were GA related.
4) **Sampling:** Given the nature of our study, we opted for a different approach compared to methodologies used in [10] or [11]. We chose not to select the top-rated posts since many relevant troubleshooting questions usually receive few votes. To ensure a democratic sampling of our data, we decided on a random sample.
   We calculated the necessary sample size n for estimating proportions within a finite population, using the established formula as described by [12]. We selected a confidence level of 95%, corresponding to an error e of 0.05 and a z-value of 1.96. The assumed value of p was set at 0.5. As a result, our data sample has 340 posts.
5) **Coding:** In this phase, we prep ared the data for analysis by converting the content from HTML to plain text using BeautifulSoup. We removed non-textual elements such as code snippets and images to focus on the textual data. The text from titles and bodies was then segmented into sentences, obtaining 3176 sentences from our data sample.

### 2) Manual classification

Our taxonomy development was an iterative and dynamic process aimed at classifying StackOverflow (SO) posts according to the facets of information needs of developers using GitHub Actions. We based our initial taxonomy on existing literature about developers’ needs and relevant information in API usage contexts [10], [11], leveraging these validated classifications to ensure a solid foundation for our study.

The first step involved identifying sentences that contained relevant information pertaining to the developer’s question. Sentences lacking relevant information were discarded. The remaining sentences, which contained relevant information, were then classified into types of Relevant Information (RI).

We employed a hybrid card sorting method, as detailed in [Ref: Zimmermann], for the classification of RIs. This method involved using existing taxonomies from previous research to kick-start our cyclical classification process. Each sentence was reviewed to determine its fit within the pre-defined categories from these taxonomies. If a sentence did not align with any existing category, a new category was created. This iterative process involved continuously updating the definitions of each category as new sentences were analyzed, ensuring a comprehensive and flexible taxonomy.

Some sentences contained information relevant to more than one type of RI. In such cases, the sentence was classified under all applicable RI types to ensure that all pertinent aspects of the information were captured accurately, reflecting the multifaceted nature of developer queries.

Once the RI categories were fully defined, each RI type was assigned to a corresponding Information Need (IN). Each IN class was clearly articulated, acknowledging that the relationship between RIs and INs is not always one-to-one; multiple RIs can correspond to a single IN. This thorough assignment process was essential for accurately mapping out the diverse information needs of developers.

The classification process was meticulously carried out by two of the co-authors, ensuring consistency and accuracy. By systematically reviewing and categorizing each sentence, we ensured that the relevant information was precisely captured and classified. This robust foundation facilitated further analysis and provided deep insights into the types of information sought by users discussing GitHub Actions.

## III. Analysis & Results

### A. RQ1: Level of interest

The monthly number of questions posted by developers on StackOverflow is illustrated in Figure 1. We utilized the Stack Exchange Data Explorer (SEDE) to access these posts, with the queries being publicly available.

Our results highlight two distinct periods: between 2019 and early 2023, and from mid-2023 onwards. In the first period, we observed a decline in the number of posts for then-popular tools like Jenkins and Azure DevOps, while GitHub Actions, released on November 13th, 2019, surged in relevance, eventually matching or surpassing these tools. This rapid growth in popularity among the developer community corroborates the findings from [13], which identified GitHub Actions as the most used tool in Continuous Integration. Notably, the number of questions for GitHub Actions reached over 350 per month.

In the second period, starting from mid-2023, there was a general decline in the number of questions across all tools, including GitHub Actions, which dropped to about 200 questions per month. This trend is not unique to GitHub Actions; similar pattern were observed for other CI tools. This decline could be attributed to the increasing popularity of large language model-based tools, such as ChatGPT. Therefore, despite the reduction in the number of questions about GitHub Actions, this could be a global effect of new querying tools rather than a decrease in interest in GitHub Actions itself.

### B. RQ2: Types of Information Needs

As a result of the manual classification process, we developed a comprehensive taxonomy for Relevant Information (RI) and Information Needs (IN), each with their respective definitions. The taxonomy consists of a total of 24 types of RIs and 8 classes of INs. The detailed taxonomy of RIs and their definitions is presented in Table [Fig: RIdef].

As specified earlier, certain RI categories serve as indicators of the presence of an IN. Table 2 provides a detailed overview of the INs, their definitions, and the associated RIs.

In some sentences, one or more types of RIs may be present. Given that each post is composed of multiple sentences, it is possible for a single post to exhibit one or more INs. This relationship between sentences, RIs, and INs highlights the complexity of developer queries and the multifaceted nature of the information they seek.

To illustrate this, we present Figure [Fig: RIPosts], which shows the number of sentences and posts, along with the percentage distribution of each type of RI and IN.

This analysis reveals that developer posts often contain multiple RIs, indicating that their information needs are diverse and complex. By understanding the distribution and association of RIs and INs, we can gain valuable insights into the types of information developers are seeking and the challenges they face when using GitHub Actions. This taxonomy not only helps in categorizing developer queries more accurately but also provides a structured approach to identifying and addressing their information needs.

## IV. Discussion

## V. Related Work

## VI. Threats to Validity & Limitations

## VII. Conclusion

## Acknowledgment

## References

[1] M. Golzadeh, A. Decan, and T. Mens, “On the rise and fall of CI services in GitHub,” in *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2022, pp. 662–672.

[2] M. Fowler and M. Foemmel, “Continuous integration,” 2006.

[3] J. Humble and D. Farley, *Continuous delivery: reliable software releases through build, test, and deployment automation*. Pearson Education, 2010.

[4] B. Fitzgerald and K.-J. Stol, “Continuous software engineering: A roadmap and agenda,” *Journal of Systems and Software*, vol. 123, pp. 176–189, 2017.

[5] Y. Zhang, Y. Wu, T. Chen, T. Wang, H. Liu, and H. Wang, “How do developers talk about GitHub actions? evidence from online software development community,” in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 2024, pp. 1–13.

[6] A. J. Ko, R. DeLine, and G. Venolia, “Information needs in collocated software development teams,” in *29th International Conference on Software Engineering (ICSE’07)*. IEEE, 2007, pp. 344–353.

[7] R. P. Buse and T. Zimmermann, “Information needs for software development analytics,” in *2012 34th International Conference on Software Engineering (ICSE)*. IEEE, 2012, pp. 987–996.

[8] A. Ouni, I. Saidani, E. Alomar, and M. W. Mkaouer, “An empirical study on continuous integration trends, topics and challenges in stack overflow,” in *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, 2023, pp. 141–151.

[9] A. Rahman, A. Partho, P. Morrison, and L. Williams, “What questions do programmers ask about configuration as code?” in *Proceedings of the 4th International Workshop on Rapid Continuous Software Engineering*, 2018, pp. 16–22.

[10] M. Liu, X. Peng, A. Marcus, S. Xing, C. Treude, and C. Zhao, “Api-related developer information needs in stack overflow,” *IEEE Transactions on Software Engineering*, vol. 48, no. 11, pp. 4485–4500, 2022.

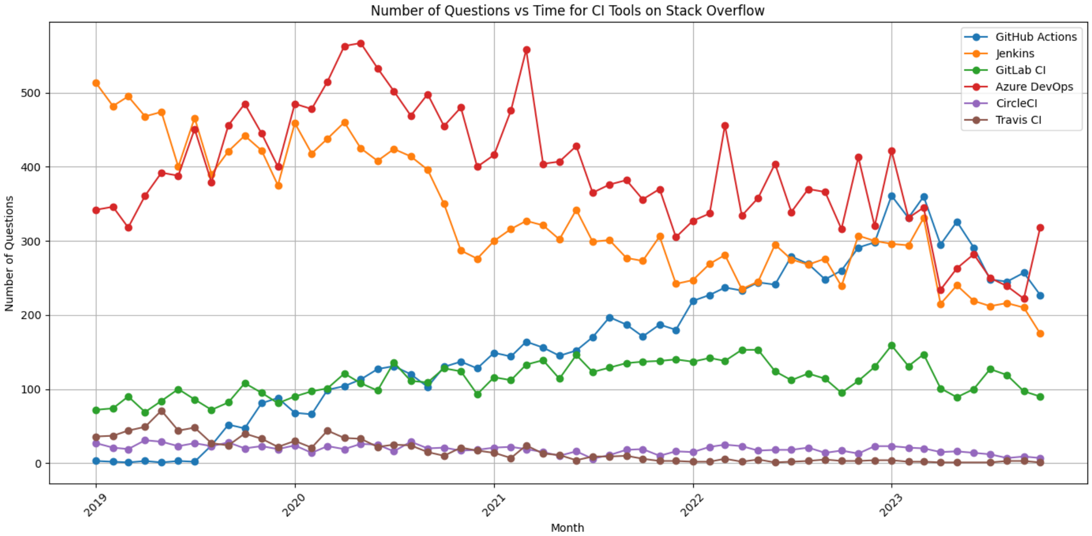

Fig. 1: Monthly number of questions with ‘github-actions’, ‘jenkins’, ‘gitlab-ci’, ‘azure-devops’, ‘circleci’, and ‘travis-ci’ tags correspondly in Stack Overflow. PV ►*probando tamaño*◄

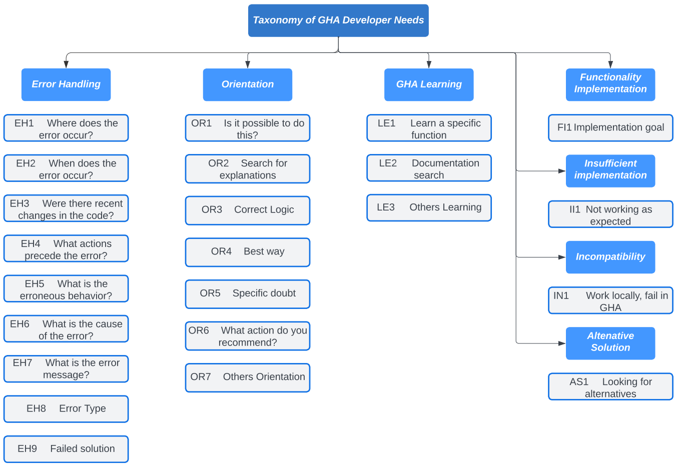
Taxonomy of GHA Developer Needs
Error Handling
EH1 Where does the error occur?
EH2 When does the error occur?
EH3 Were there recent changes in the code?
EH4 What actions precede the error?
EH5 What is the erroneous behavior?
EH6 What is the cause of the error?
EH7 What is the error message?
EH8 Error Type
EH9 Failed solution
Orientation
OR1 Is it possible to do this?
OR2 Search for explanations
OR3 Correct Logic
OR4 Best way
OR5 Specific doubt
OR6 What action do you recommend?
OR7 Others Orientation
GHA Learning
LE1 Learn a specific function
LE2 Documentation search
LE3 Others Learning
Functionality Implementation
FI1 Implementation goal
Insufficient implementation
II1 Not working as expected
Incompatibility
IN1 Work locally, fail in GHA
Altenative Solution
AS1 Looking for alternatives

Fig. 2: Information Needs (IN), their Definitions, and Associated Relevant Information (RI)

[11] S. Beyer, C. Macho, M. Di Penta, and M. Pinzger, “Automatically classifying posts into question categories on stack overflow,” in *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*, 2018, pp. 211–21 110.

[12] M. Triola, W. Goodman, G. LaBute, R. Law, and L. MacKay, *Elementary Statistics*. Pearson Education Canada, 2009. [Online]. Available: https://books.google.ch/books?id=qZIhPwAACAAJ

[13] T. Blog. (2023) Best continuous integration tools for 2023 – survey results. [Online]. Available: https://blog.jetbrains.com/teamcity/2023/07/best-ci-tools/