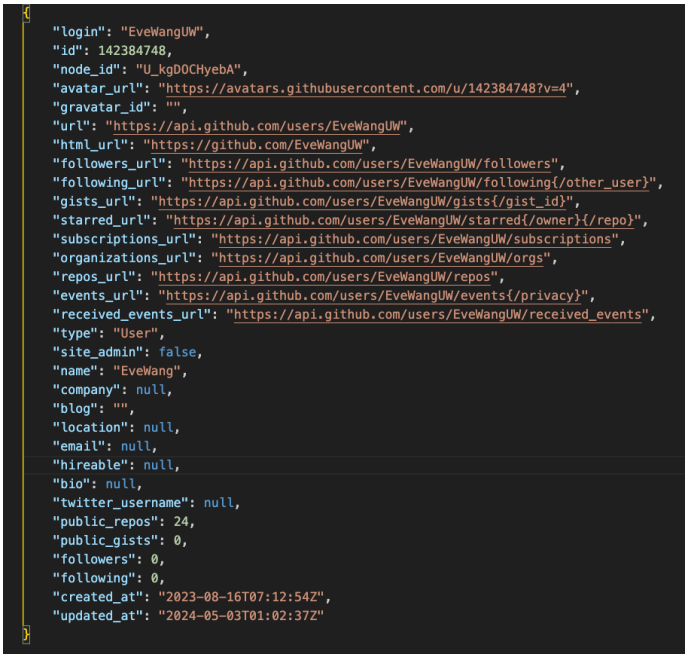
G7 Assignment

A sample of the existing information structures:

1. Information structures of ‘user’:



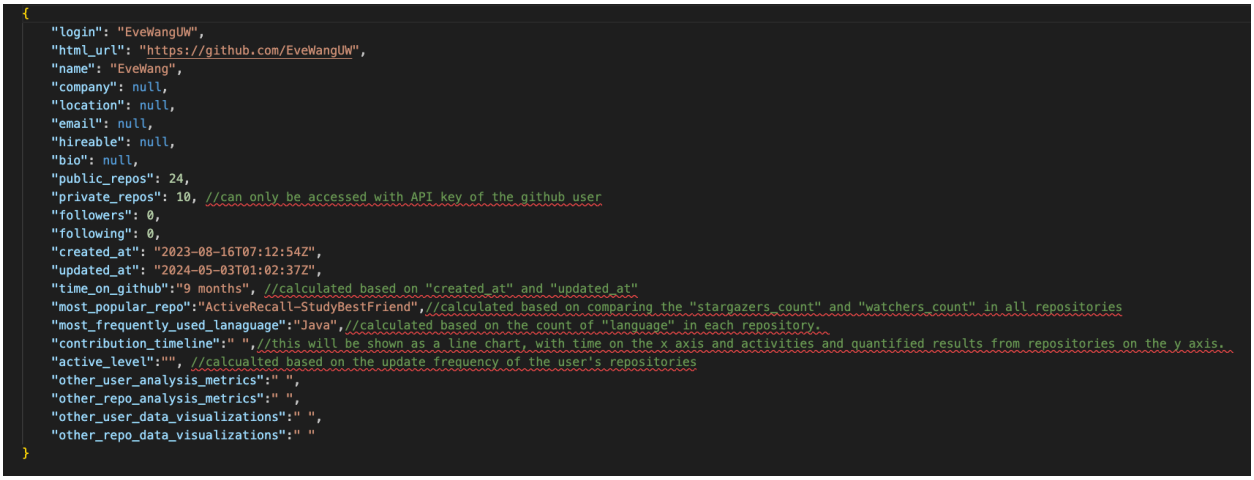

```
{
    "login": "EveWangUW",
    "id": 142384748,
    "node_id": "U_kgDOCHyebA",
    "avatar_url": "https://avatars.githubusercontent.com/u/142384748?v=4",
    "gravatar_id": "",
    "url": "https://api.github.com/users/EveWangUW",
    "html_url": "https://github.com/EveWangUW",
    "followers_url": "https://api.github.com/users/EveWangUW/followers",
    "following_url": "https://api.github.com/users/EveWangUW/following{/other_user}",
    "gists_url": "https://api.github.com/users/EveWangUW/gists{/gist_id}",
    "starred_url": "https://api.github.com/users/EveWangUW/starred{/owner}{/repo}",
    "subscriptions_url": "https://api.github.com/users/EveWangUW/subscriptions",
    "organizations_url": "https://api.github.com/users/EveWangUW/orgs",
    "repos_url": "https://api.github.com/users/EveWangUW/repos",
    "events_url": "https://api.github.com/users/EveWangUW/events{/privacy}",
    "received_events_url": "https://api.github.com/users/EveWangUW/received_events",
    "type": "User",
    "site_admin": false,
    "name": "EveWang",
    "company": null,
    "blog": "",
    "location": null,
    "email": null,
    "hireable": null,
    "bio": null,
    "twitter_username": null,
    "public_repos": 24,
    "public_gists": 0,
    "followers": 0,
    "following": 0,
    "created_at": "2023-08-16T07:12:54Z",
    "updated_at": "2024-05-03T01:02:37Z"
}
```

2. Information structures of ‘repository’:

```
{
    "id": 793294266,
    "node_id": "R_kgDOL0i1ug",
    "name": "ActiveRecall-StudyBestFriend",
    "full_name": "EveWangUW/ActiveRecall-StudyBestFriend",
    "private": false,
    "owner": {
        "login": "EveWangUW",
        "id": 142384748,
        "node_id": "U_kgDOCHyebA",
        "avatar_url": "https://avatars.githubusercontent.com/u/142384748?v=4",
        "gravatar_id": "",
        "url": "https://api.github.com/users/EveWangUW",
        "html_url": "https://github.com/EveWangUW",
        "followers_url": "https://api.github.com/users/EveWangUW/followers",
        "following_url": "https://api.github.com/users/EveWangUW/following{/other_user}",
        "gists_url": "https://api.github.com/users/EveWangUW/gists{/gist_id}",
        "starred_url": "https://api.github.com/users/EveWangUW/starred{/owner}{/repo}",
        "subscriptions_url": "https://api.github.com/users/EveWangUW/subscriptions",
        "organizations_url": "https://api.github.com/users/EveWangUW/orgs",
        "repos_url": "https://api.github.com/users/EveWangUW/repos",
        "events_url": "https://api.github.com/users/EveWangUW/events{/privacy}",
        "received_events_url": "https://api.github.com/users/EveWangUW/received_events",
        "type": "User",
        "site_admin": false
    },
    "html_url": "https://github.com/EveWangUW/ActiveRecall-StudyBestFriend",
    "description": null,
    "fork": false,
    "url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend",
    "forks_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/forks",
    "keys_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/keys{/key_id}",
    "collaborators_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/collaborators{/collaborator}",
    "teams_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/teams",
    "hooks_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/hooks",
    "issue_events_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/issues/events{/number}",
    "events_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/events",
    "assignees_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/assignees{/user}",
    "branches_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/branches{/branch}",
    "tags_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/tags",
    "blobs_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/git/blobs{/sha}",
    "git_tags_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/git/tags{/sha}",
    "git_refs_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/git/refs{/sha}",
    "trees_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/git/trees{/sha}",
    "statuses_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/statuses/{sha}",
    "languages_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/languages",
    "stargazers_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/stargazers",
    "contributors_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/contributors",
    "subscribers_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/subscribers",
    "subscription_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/subscription",
    "commits_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/commits{/sha}",
    "git_commits_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/git/commits{/sha}",
    "comments_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/comments{/number}",
    "issue_comment_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/issues/comments{/number}",
    "contents_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/contents/{+path}",
    "compare_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/compare/{base}...{head}",
    "merges_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/merges",
    "archive_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/{archive_format}{/ref}",
    "downloads_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/downloads",
    "issues_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/issues{/number}",
    "pulls_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/pulls{/number}",
    "milestones_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/milestones{/number}",
    "notifications_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/notifications{?since,all,participating}",
    "labels_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/labels{/name}",
    "releases_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/releases{/id}",
    "deployments_url": "https://api.github.com/repos/EveWangUW/ActiveRecall-StudyBestFriend/deployments",
    "created_at": "2024-04-28T23:51:14Z",
    "updated_at": "2024-04-28T23:52:01Z",
    "pushed_at": "2024-04-28T23:51:24Z",
    "git_url": "git://github.com/EveWangUW/ActiveRecall-StudyBestFriend.git",
    "ssh_url": "git@github.com:EveWangUW/ActiveRecall-StudyBestFriend.git",
    "clone_url": "https://github.com/EveWangUW/ActiveRecall-StudyBestFriend.git",
    "svn_url": "https://github.com/EveWangUW/ActiveRecall-StudyBestFriend",
    "homepage": null,
    "size": 358,
    "stargazers_count": 0,
    "watchers_count": 0,
    "language": "JavaScript",
    "has_issues": true,
    "has_projects": true,
    "has_downloads": true,
    "has_wiki": true,
    "has_pages": false,
    "has_discussions": false,
    "forks_count": 0,
    "mirror_url": null,
    "archived": false,
    "disabled": false,
    "open_issues_count": 0,
    "license": null,
    "allow_forking": true,
    "is_template": false,
    "web_commit_signoff_required": false,
    "topics": [],
    "visibility": "public",
    "forks": 0,
    "open_issues": 0,
    "watchers": 0,
    "default_branch": "main",
    "permissions": {
        "admin": true,
        "maintain": true,
        "push": true,
        "triage": true,
        "pull": true
    }
}
```

A sample of the new structure (combining and manipulating some attributes from 'user' and some attributes from 'repository')

Note: the green comments are shown here for explanation purposes. To conform to the JSON format, I have removed these comments in the uploaded json files.



```
{
    "login": "EveWangUW",
    "html_url": "https://github.com/EveWangUW",
    "name": "EveWang",
    "company": null,
    "location": null,
    "email": null,
    "hireable": null,
    "bio": null,
    "public_repos": 24,
    "private_repos": 10, //can only be accessed with API key of the github user
    "followers": 0,
    "following": 0,
    "created_at": "2023-08-16T07:12:54Z",
    "updated_at": "2024-05-03T01:02:37Z",
    "time_on_github":"9 months", //calculated based on "created_at" and "updated_at"
    "most_popular_repo":"ActiveRecall-StudyBestFriend",//calculated based on comparing the "stargazers_count" and "watchers_count" in all repositories
    "most_frequently_used_lanaguage":"Java",//calculated based on the count of "language" in each repository.
    "contribution_timeline":" ",//this will be shown as a line chart, with time on the x axis and activities and quantified results from repositories on the y axis.
    "active_level":"", //calcualted based on the update frequency of the user's repositories
    "other_user_analysis_metrics":" ",
    "other_repo_analysis_metrics":" ",
    "other_user_data_visualizations":" ",
    "other_repo_data_visualizations":" "
}
```

My project will use the data from GitHub API endpoints and restructure them to provide insights on the GitHub user, such as their most frequently used language, what is their level of activeness on GitHub, etc.

Q1: Data Quality:

1. For the specific data scope, I will use variables on repos such as 'name', 'description', 'stargazers_count', 'language' and users such as 'login'(username), 'name'(real name), 'followers', 'company'.

This can help realize the **consistency of the data** because we can trust that these attributes are uniformly used across different users and repositories. At the same time, some calculated metrics, such as the "most_popular_repo", "most_frequently_used_lanaguage", "contribution_timeline", "active_level", etc., may need to be modified due to the change in business needs, and we need to communicate with **stakeholders** in a timely manner, and on the basis of reaching a consensus, modify the **data generation plan** or data scheme to update their calculation method so as to achieve **consistency** (Dilmegani, 2024). In this way, may information structure can promote maintaining and measuring quality easier.

2. For the data format, I will retrieve the data from the GitHub API endpoints in JSON format, which is the type of structure I will be using.

We can validate the JSON responses against predefined schemas or **data generation plans** to help us find out if our data retrieved have deviations from expectations and update our standards in a timely manner according to the new data standards, enabling data **calibration**. We can also have a schedule for making API calls, ensuring that data is **timely** and **calibrated** (McCord et al., 2022). This can help my information structure promote maintaining and measuring quality in an easier manner.

Q2: Information Security:

1. For the specific data scope, I will use variables such as 'private' (return true if it is a private repository). This can help with **data classification** and give repositories different levels of security protection, e.g., it will only allow access to the private repository by providing the API key of the repository owner. We can also establish the **encryption for management strategy** in this project, create encryption for private repository data, and define the encryption standards for handling these private and sensitive data. This can help ensure that my information structure can enable any information security issues to be identified and managed in the future.

2. For the data access of my data, I will utilize APIs to fetch data from the GitHub API, which will be my access method. We can **back up the data** from the previous API call results and store them in a database. We can also show the updated timelines of a repository and recover our application if any information is attacked with these backed-up data. Moreover, currently, the bearer token or API key of the GitHub user should be passed in the GET request when retrieving their private repository data from the GitHub APIs. Building on top of this, we can also add other **encryption** strategies and policies, such as adding a check on whether it is a real person making the request and further encrypting the bearer token to enhance information security. We can also, in the process of the user requesting information, enhance the monitoring of attacks, attempts to decrypt, and other undesirable acts, which can help identify and manage any information security issues that may occur in the future (Exabeam, 2024).

References:

Dilmegani, C. (2024). *Data Quality Assurance: Importance & Best Practices in 2024.* AIMultiple Research. https://research.aimultiple.com/data-quality-assurance/

Exabeam. (2024). *The 12 Elements of an Information Security Policy*. exabeam. https://www.exabeam.com/explainers/information-security/the-12-elements-of-an-information-security-policy/

McCord, S. E., et al. (2022). *Ten Practical Questions to improve data quality*, Rangelands,

Volume 44, Issue 1, 2022, Pages 17-28, ISSN 0190-0528, https://doi.org/10.1016/j.rala.2021.07.006. (https://www.sciencedirect.com/science/article/pii/S0190052821000699)