SemanGit: A Linked Dataset from git

Dennis Oliver Kubitzka, Matthias Böckmann & Damien Graux

<http://www.semangit.de/>

<https://github.com/SemanGit/SemanGit>

Dennis Oliver Kubitzka

29th October 2019

ISWC 2019 - Auckland New Zealand
Content

• Motivation
• What is SemanGit
• First Analyses to showcase Potential
• Sustainability Plan
• Current Challenges and Future Work
Motivation
Motivation

“In other words, Microsoft is not paying $7.5 billion for GitHub for its ability to make money (its financial value). It’s paying for the access it gets to the legions of developers who use GitHub’s code repository products on a daily basis” [...] 

Paul V. Weinstein: “Why Microsoft is willing to pay so much for GitHub.”, 

*Harvard Business Review*, 06 June 2018
The Data of GitHub\textsuperscript{[1]}

- 31 M+ developers
- Following each other
- Contributing to the same projects
- Referencing their homepages, Social Media, Employers etc.
The Data of GitHub\textsuperscript{[1]}

- 31 M+ developers
  - Following each other
  - Contributing to the same projects
  - Referencing their homepages, Social Media, Employers etc.

- 2 M+ organizations
  - Institutional, Business, Non-Profit
  - With an evolving set of members
The Data of GitHub\textsuperscript{[1]}

- 31 M+ developers
  - Following each other
  - Contributing to the same projects
  - Referencing their homepages, Social Media, Employers etc.

- 2 M+ organizations
  - Institutional, Business, Non-Profit
  - With an evolving set of members

- 96 M+ repositories
  - Information about Watchers, Starring, Issues ...
How to Access this data?

1. Be GitHub

2. Acquire GitHub

3. Use the API with a rate limit of 5,000 queries per hour\textsuperscript{[2]} and wait several years
How to Access this data?

1. Be GitHub

2. Acquire GitHub

3. Use the API with a rate limit of 5,000 queries per hour[2] and wait several years
   Use different datasets, listening to or scraping GitHub.
Datasets

• Multiple Providers for GitHub Data
  • GHTorrent, GHArchive, Public Git Archive [...]

Datasets

• Multiple Providers for GitHub Data
  • GHTorrent, GHArchive, Public Git Archive [...] 

• Each following their own approach
• Each focusing on different subsets of the whole available data
• Each providing differently structured Datasets
Datasets

- Multiple Providers for GitHub Data
  - GHTorrent, GHArchive, Public Git Archive [...]
- Each following their own approach
- Each focusing on different subsets of the whole available data
- Each providing differently structured Datasets

Not only for GitHub, but also for GitLab, SourceForge etc.
SemanGit

**Problem:** Multiple partially overlapping datasets following different Design Decisions, covering different Subsets of Data

**Solution:** Use Semantic Technologies to create a Knowledge Graph.
SemanGit = git + Semantics

- Semantic Strengths
  - Data from multiple Providers and APIs
  - Enriching by linking with LOD datasets

- Graph Strengths
  - Enabling Graph traversals for evaluating local structures
SemanGit Ontology

Ontology available at [www.SemanGit.de](http://www.SemanGit.de) and GitHub

Visualisation with WebVOWL [3]
SemanGit Ontology

Multi-Layered by Degrees of Abstraction

- Git Core
- Abstraction of Social Features
- Provider Specific Classes

Ontology available at www.SemanGit.de and GitHub
SemanGit Ontology

Multi-Layered by Degrees of Abstraction
- Git Core
- Abstraction of Social Features
- Provider Specific Classes

Structured by Inheritance

Ontology available at [www.SemanGit.de](http://www.SemanGit.de) and GitHub
SemanGit Dataset

Contains for now:

- Over **22 billion triples**, with high interlinkage between classes.
- Links from GitHub to DBpedia in terms of Geolocations.

Documentation, Scripts and Dataset available at [www.SemanGit.de](http://www.SemanGit.de) and GitHub
SemanGit Dataset

Contains for now:

- Over 22 billion triples, with high interlinkage between classes.
- Links from GitHub to DBpedia in terms of Geolocations.
- Information about:
  - Social Interactions between Users ..
  - Reported Code Issues …
  - Organisations and Projects and the respective follower and starrers
  - Commit Structure within Projects incl. Pull Requests, used Programming Languages
  - …

Documentation, Scripts and Dataset available at [www.SemanGit.de](http://www.SemanGit.de) and GitHub
Figure 1: Comparison of used Programming Languages for China and the United States of America
Figure 2: Colored Map of high vice-versa Cooperation with Indonesia
The shades of blue encode which nation has a high index. Nations with less than 1000 users are omitted in grey.
Figure 3: Density Plot of the count of internal Follow Relations within an Organisation
Grouping by the number of organization members and the success of the Organisation
Sustainability Plan

- We guarantee our resources to be available at least until 2022
- Newest Dataset and Ontologies are published at our Website
- Old Dumps can be recreated by using Scripts published on GitHub

Everything is Open-Source!
Current Challenges and Future Work

• Include new Data Sources like GitLab
• Query APIs for missing Data
• Add Interlinkage to other Datasets for:
  • Scientific Publication
  • Official Organisations and Companies
  • (eventually Social Media Platforms)
• Provide public SparQL Endpoint
References


Thank You ... Questions?

dennis.oliver.kubitza@iais.fraunhofer.de

http://www.semangit.de/