Towards Semantically Structuring GitHub

Dennis Oliver Kubitza, Matthias Böckmann & Damien Graux

<http://www.semangit.de/>   <https://github.com/SemanGit/SemanGit>

Challenges with VCS Data

- Multiple Hosts offering remote git Infrastructure (GitHub, Gitlab, SourceForge, …)
  - … implementing different non-standard features
  - … offering Data Access with limited APIs

- Multiple existing attempts to collect **partial** data from GitHub etc.
  - E.g. for some time interval or only about certain event types
  - Each applying their own data model
  - No links between those heterogeneous sources

Achieved Goals

- Develop a publicly available ontology
  - Modelling knowledge about the git protocol
  - Extensible by provider specific features
  - Tailored to capture all available data from public databases

Future Work

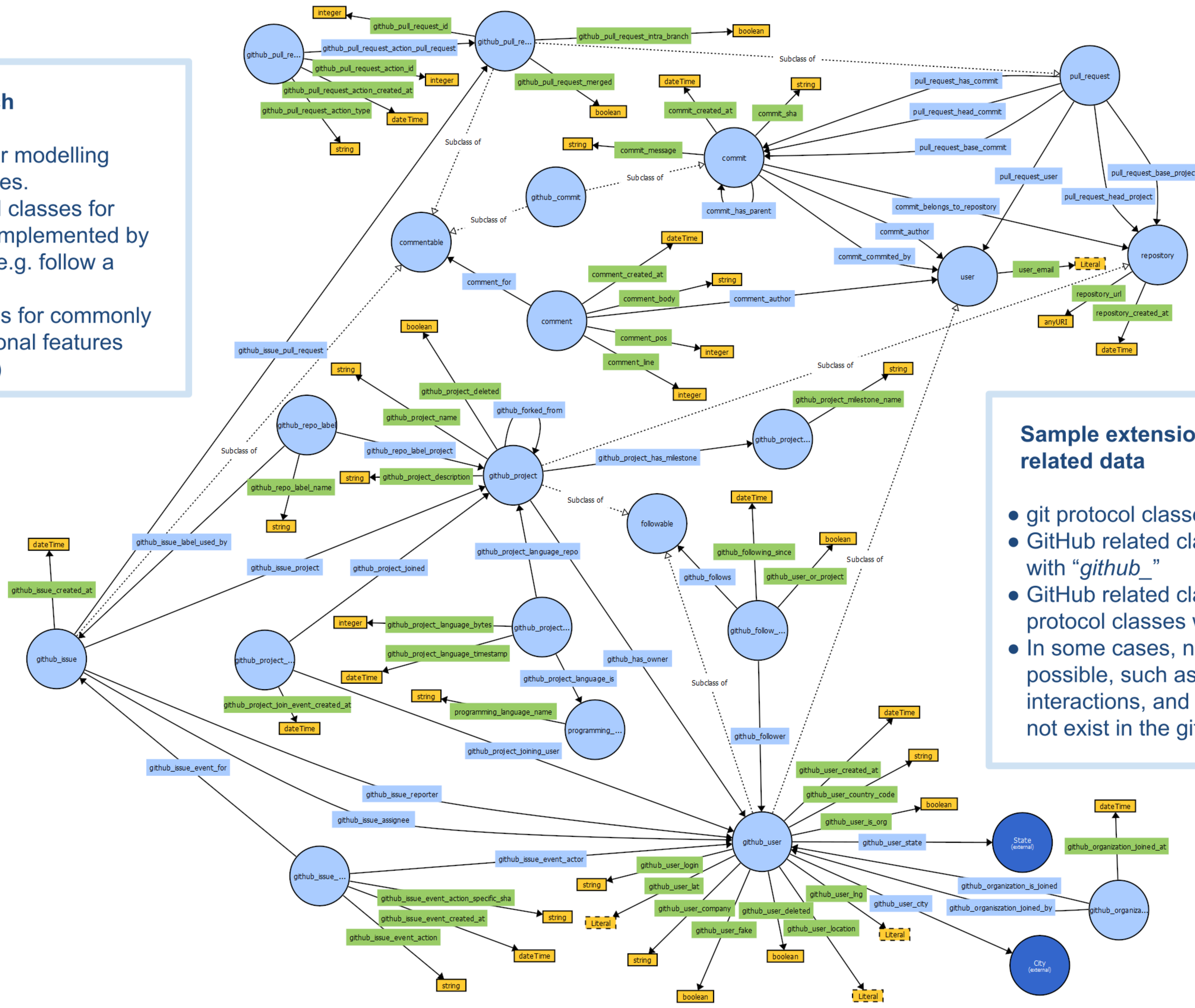
- Generalise and extend ontology for arbitrary sources
  - Using git glossary as naming reference
  - Using provider's API for completeness

- Extend usage of LOD vocabularies to allow interlinkage

The Semantic Git Ontology

Modelling Approach

- Contains classes for modelling git core-functionalities.
- Introduce additional classes for capturing features implemented by multiple providers (e.g. follow a user)
- Intermediate classes for commonly implemented additional features (e.g. issue tracking)



Sample extension to capture GitHub related data

- git protocol classes have no prefix
- GitHub related classes are prefixed with "*github_*"
- GitHub related classes inherit from git protocol classes where applicable
- In some cases, no inheritance is possible, such as issue tracking, social interactions, and comments, which do not exist in the git protocol

Visualised with WebVOWL
http://vowl.visualdataweb.org/webvowl.html

Further Reading

SemanGit: A Linked Dataset from git *by* Dennis Oliver Kubitza, Matthias Böckmann & Damien Graux *in* ISWC, 2019.

Acknowledgements

This work is partly supported by the German Federal Ministry of Education and Research (BMBF) in the context of the research project "Industrial Data Space Plus" (GA 01IS17031) as well as the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies" (CCIT); by the EU H2020 project "QualiChain" (GA 822404); and by the ADAPT Centre for Digital Content Technology funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

Fraunhofer IAIS

SMART DATA ANALYTICS FROM DATA TO KNOWLEDGE

UNIVERSITÄT BONN

Engaging Content Engaging People