SemanGit

Final Presentation

Matthias Böckmann, Dennis Kubitza

Lab 4314

18. Oktober 2018

Contents

1. Theory
   - Introduction
   - Problem
   - Proposed Solutions
   - Implementation and Evaluation
2. Application
   - Use Cases
   - Lessons Learned
3. Conclusions and Future Work

What is SemanGit ?

Git A Protocol for file version control [1]

GitHub A Hoster/Web-Storage with and underlying implementation of Git [2]

SemanGit A semantic Database containing

- an Ontology for the Git Protocol and GitHub
- a huge Dataset extracted from GitHub

Motivation

- Make Meta-Data from Git more accessible. (for Users)

Main Goal

- Provide a complete Dataset which is still usable in practice.

Main Problem

1. 28 million users [2]
   + 57 million repositories [2]
   + 1200 million commits [4]
   = Hundreds of GiB of Data to Extract and Convert.

2. Git + GitHub + (GitLab + BitBucket + ...)+ GHTorrent = Ontology ???

Proposed Solutions - Data

- Extract Data from GitHub API

Proposed Solutions - Data

- Extract Data from GitHub API

Proposed Solutions - Data

- Extract Data from GitHub API - **BAD IDEA** (5000 Queries / hour [3])

Proposed Solutions - Data

- Extract Data from GitHub API - **BAD IDEA** (5000 Queries / hour [3])
- Extract Data from GHTorrent [7]

Proposed Solutions - Data

- Extract Data from GitHub API - **BAD IDEA** (5000 Queries / hour [3])
- Extract Data from GHTorrent [7]
- Try to keep RDF-Expansion factor as small as possible (Turtle)

Proposed Solutions - Data

- Extract Data from GitHub API - **BAD IDEA** (5000 Queries / hour [3])
- Extract Data from GHTorrent [7]
- Try to keep RDF-Expansion factor as small as possible (Turtle)
- Ensure fault tolerant, autonomous processing for extraction.

Proposed Solutions - Ontology

- Strictly Distinguish between Git and additional features

Proposed Solutions - Ontology

- Strictly Distinguish between Git and additional features
- Ensure re-usability by using Class Hierarchies

Proposed Solutions - Ontology

- Strictly Distinguish between Git and additional features
- Ensure re-usability by using Class Hierarchies
- Ensure Completeness over available Data

Proposed Solutions - Ontology

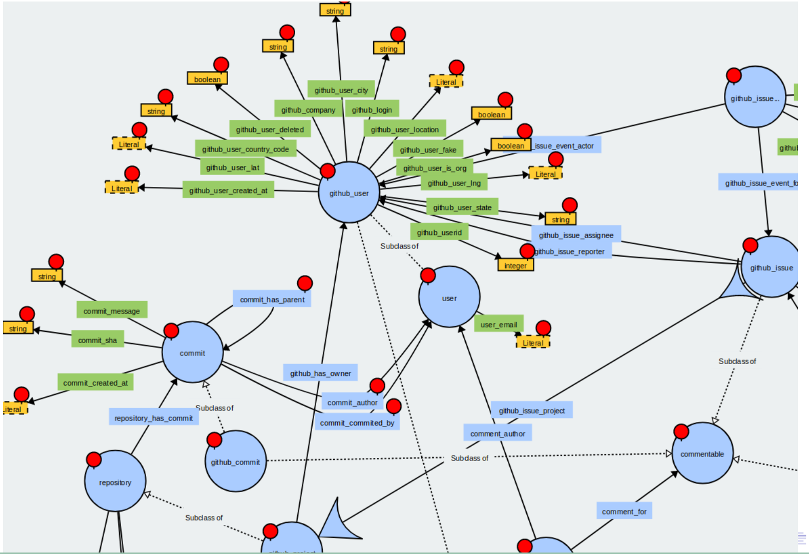
- Strictly Distinguish between Git and additional features
- Ensure re-usability by using Class Hierarchies
- Ensure Completeness over available Data

Proposed Solutions - Ontology

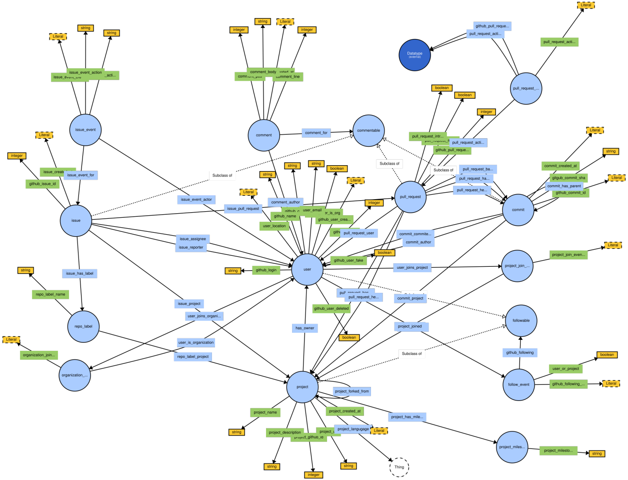
- Strictly Distinguish between Git and additional features
- Ensure re-usability by using Class Hierarchies
- Ensure Completeness over available Data

Ontology: [5]

Proposed Solutions - Ontology



Proposed Solutions - Ontology Example



Proposed Solutions - Ontology Summary

Our final Ontology has:

Classes	: 22
Properties	: 80

Output Compression

- Smart Prefixes

Output Compression

- Smart Prefixes
  - Prefixes a-zA-Z (Base 52)

Output Compression

- Smart Prefixes
  - Prefixes a-zA-Z (Base 52)
  - Empty Prefix ":"

Output Compression

- Smart Prefixes
  - Prefixes a-zA-Z (Base 52)
  - Empty Prefix ":"
  - $< http : //semangit.de/ontology/ghuser\#1327 > \rightarrow b : 1327$

Output Compression

- Smart Prefixes
  - Prefixes a-zA-Z (Base 52)
  - Empty Prefix ":"
  - $< http : //semangit.de/ontology/ghuser\#1327 > \rightarrow b : 1327$
- Integer Representation

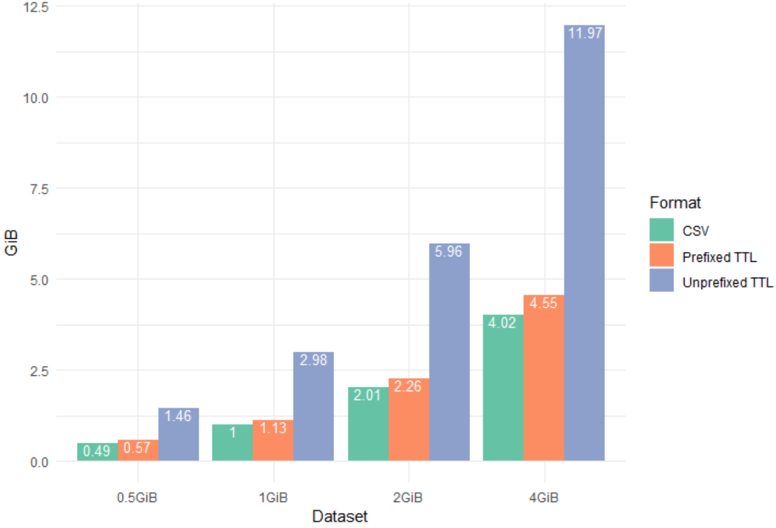
Output Compression

- Smart Prefixes
  - Prefixes a-zA-Z (Base 52)
  - Empty Prefix ":"
  - $< http : //semangit.de/ontology/ghuser\#1327 > \rightarrow b : 1327$
- Integer Representation
  - Base64 approach

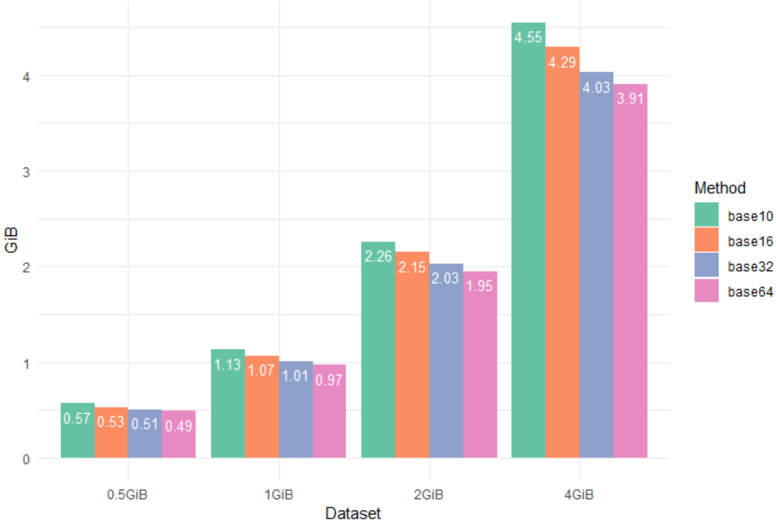
Output Compression

- Smart Prefixes
  - Prefixes a-zA-Z (Base 52)
  - Empty Prefix ":"
  - $< http://semangit.de/ontology/ghuser\#1327 > \rightarrow b:1327$
- Integer Representation
  - Base64 approach
  - $b:1327 \rightarrow b:l5$

Evaluation


12.5
10.0
7.5
5.0
2.5
0.0
GiB
0.49
0.57
1.46
1
1.13
2.98
2.01
2.26
5.96
4.02
4.55
11.97
0.5GiB
1GiB
2GiB
4GiB
Dataset
Format
CSV
Prefixed TTL
Unprefixed TTL

Evaluation


GiB
0
1
2
3
4
0.57
0.53
0.51
0.49
1.13
1.07
1.01
0.97
2.26
2.15
2.03
1.95
4.55
4.29
4.03
3.91
0.5GiB
1GiB
2GiB
4GiB
Dataset
Method
base10
base16
base32
base64

Evaluation

Tab.: Absolute Conversion Run-Time

No-Prefix	base10	base16	base32	base64
413	216	228	237	244

Runtime in seconds, 4GiB of Data

AMD Athlon X4 860K (4x4.0GHZ), 16GB DDR3-1866, 2TB HDD

Use Cases

- Headhuntress

Use Cases

- Headhuntress
- Check for Plagiarism

Use Cases

- Headhuntress
- Check for Plagiarism
- Economists

Lessons Learned

- Structural changes in GHTorrent

Lessons Learned

- Structural changes in GHTorrent
  - New Tables

Lessons Learned

- Structural changes in GHTorrent
  - New Tables
  - Columns added/removed

Lessons Learned

- Structural changes in GHTorrent
  - New Tables
  - Columns added/removed
  - Column order changed

Lessons Learned

- Structural changes in GHTorrent
  - New Tables
  - Columns added/removed
  - Column order changed
- Remember HDT? [6]

Lessons Learned

- Structural changes in GHTorrent
  - New Tables
  - Columns added/removed
  - Column order changed
- Remember HDT? [6]
  - Careful with tools that are still in research

Conclusion

We managed to implement a system which is capable of

- Extracting a Dataset of 12.6 billion triples
  - ... within just some days of computation time
  - ... consisting of every public and valid information that is provided
  - ... while even using less storage than the corresponding Database and the raw data.

Future Work

- Add more providers

Future Work

- Add more providers
- Machine Learning and SANSA Stack

Future Work

- Add more providers
- Machine Learning and SANSA Stack
- Integer Representation and UTF8?

References I

git - A free and open source distributed version control system. https://git-scm.com/about. Accessed: 18. Oktober 2018.

GitHub About. https://github.com/about. Accessed: 18. Oktober 2018.

GitHub API v3. https://developer.github.com/v3/. Accessed: 18. Oktober 2018.

SemanGit on GitHub. https://github.com/DennisKubitza/MA-INF-4314_SemanGit. Accessed: 18. Oktober 2018.

SemanGit Ontology on GitHub. https://github.com/DennisKubitza/MA-INF-4314_SemanGit/blob/master/Theory/ontology/semangitontology.ttl. Accessed: 18. Oktober 2018.

References II

Javier D. Fernandez, Miguel A. Martinez-Prieto, Claudio Gutierrez, Axel Polleres, and Mario Arias. Binary rdf representation for publication and exchange (hdt). *Web Semantics: Science, Services and Agents on the World Wide Web*, 19:22 – 41, 2013.

G. Gousios and D. Spinellis. Ghtorrent: Github's data from a firehose. In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, pages 12–21, June 2012.

Thank you for your attention

Any Questions?