Exploring (the impact of) Modeling/Documentation in Open Source Project

Software Engineering is a knowledge-intensive activity. In open source projects teams work remotely and hence depend even more than co-located teams on the explicit codification of knowledge for sharing across the team. However, since the advent of the Agile-manifesto, documentation - including modeling of software architecture and design - are frowned upon as a 'necessary evil'. Amongst programmers, code is seen as 'the only truth' and representing knowledge through UML models as an unnecessary effort.
In this hackathon we would like to better understand:

1. What role does design documentation play in open source projects?
2. Where should we look for the benefits of using design documentation ?
    - In software quality? In efficiency of software development (process)?

Over the past years, Chaudron c.s. have collected 95,000 UML diagrams from more than 24,000 GitHub projects. This has been the first effort at identifying the explicit codification of design knowledge in open source projects. A large-scale survey was subsequently conducted to discover the practices and developer's perception/motivation of using UML in OSS projects [2]. Meta-data of the projects and UML models is stored in Lindholmen database (see description below). Access to this database will be given to all participants of the Hackathon as a starting point to explore the research topics.

Some questions to get inspired:

- Is design (UML) done upfront or is it done after the fact?
- How detailed is the design documentation?
- How up to date is the design documentation?
- How does UML help with onboarding? Not necessarily more new people joining the project, but maybe getting the ones that join up to speed quicker?
- Does the quality of documentation impact the quality of the code design?
    - In terms of defects?
    - Or design quality - such as modularity?
- Does UML help decrease defect repair times?
- Are design docs more useful early in the creation stage of a project than later on, when things mature / stabilize
- Do code comments serve as an alternative to design documentation for sharing knowledge?
- People learn differently. Who on a team benefits most from design docs?

Lindholmen database

The Lindholmen database is hosted at Gothenburg and Chalmers University, Sweden and its homepage is at http://oss.models-db.com/. The database was first introduced at MODELS conference in San-Malo, France (October 2016). Papers [1] and [3] mentioned in details the steps of constructing the database. Over the time, the database has been curated and evolved in several directions:

- Classification of UML diagrams by types, i.e. '*class diagrams*', '*sequence diagrams*' and '*Other*' (added 2017).
- Extraction of content of more than 26,000 class diagrams from images (added early 2018). This data will allow researchers to study quality of UML diagrams that are used in OSS projects.
- Addition of textual documentation of the projects. This enables researchers to exploring the correlation/impacts of using (design) documents in OSS projects.

As it is now, the database contains two main parts:

- **UML Projects Meta-data** contains meta-data of the projects that use UML models. Example of the meta-data are the commits where UML models were introduced/updated, the developers who committed and the co-changed files.
- **Class Diagram Details** contains contents extracted from class diagram images. Example of the class diagram components stored in this database are class attributes, methods, method parameters, associations between classes.

Figure 1 shows the database schema of the database. Detailed description of the database schema can be found at http://oss.models-db.com/Downloads/SATToSE2018_Hackathon.

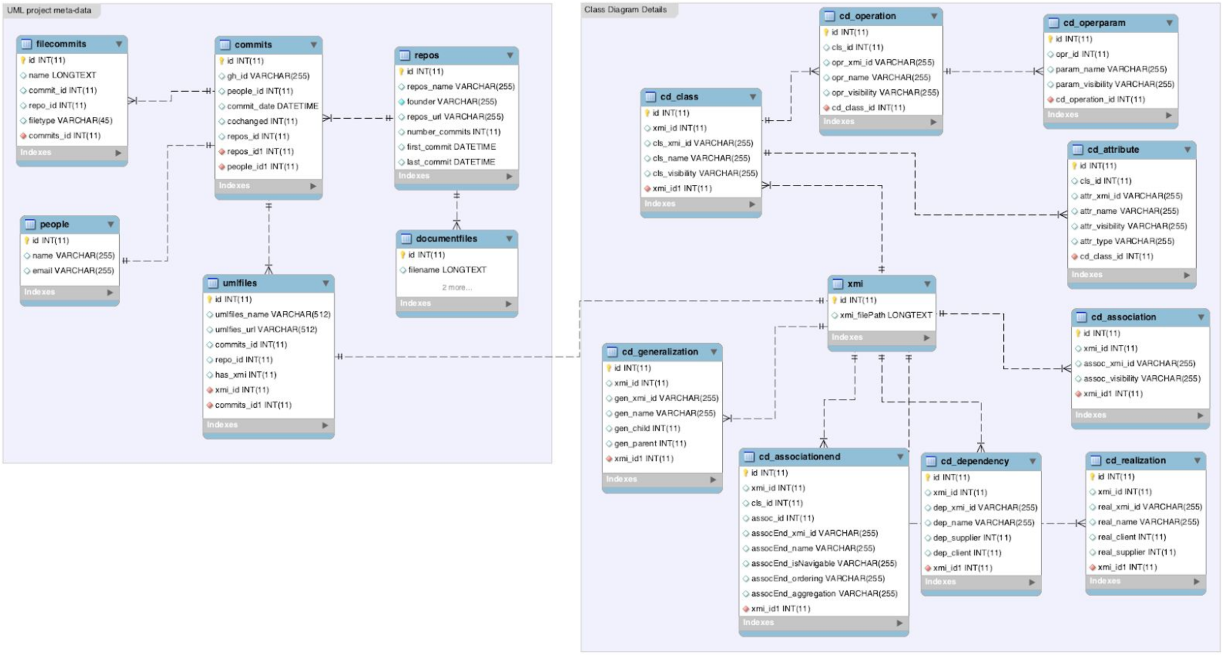

Figure 1. Database schema of Lindholmen database

In the hackathon, participants will be given access to a small set of the database (full-version will be provided on request at SATToSE). This set of database contains data of 5,000 OSS projects with more than 15,000 UML models. Size of the MySQL self-contained dump-file is about 400MB (zip) - Link to download the database will be provided at the introduction of the hackathon.

Background Papers

[1] The Quest for Open Source Projects that Use UML: Mining GitHub
http://oss.models-db.com/Downloads/models.pdf

[2] “Practices and Perceptions of UML Use in Open Source Projects”
http://oss.models-db.com/Downloads/ho-quang.icse-seip2017.pdf
Survey replication package: http://oss.models-db.com/2017-icse-seip-uml/

[3] "An extensive collection of UML files in GitHub"
https://www.computer.org/csdl/proceedings/msr/2017/1544/00/07962411.pdf

[4] “Sketches and Diagrams in Practice”
https://arxiv.org/pdf/1706.09172.pdf

[5] “Perceptions of Software Modeling: A Survey of Software Practitioners”
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.610.1440&rep=rep1&type=pdf

[6] “How Software Engineers Use Documentation: The State of the Practice”
https://www2.cs.duke.edu/courses/fall11/cps196.1/classwork/Lethbridge-Singer-Forward-2003.pdf

[7] “How effective is UML modeling?”
https://link.springer.com/content/pdf/10.1007%2Fs10270-012-0278-4.pdf

[8] “Open Source barriers to entry, revisited: A sociotechnical perspective”
ftp://ftp.cs.orst.edu/pub/burnett/icse18-OSSbarriers+gender.pdf

[9] Yates R.Y. (2014), Onboarding in Software Engineering, PhD Thesis, University of Limerick.
Her PhD thesis is available on https://ulir.ul.ie/handle/10344/4272
https://www.icse2018.org/event/icse-2018-new-ideas-and-emerging-results-dazed-measuring-the-cognitive-load-of-solving-technical-interview-problems-at-the-whiteboard