

The Quest for Open Source Projects that use UML

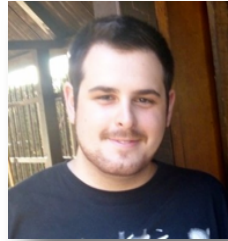
-- Mining GitHub --



Regina
Hebig *



Truong Ho-
Quang *



Miguel-Angel
Fernandez +



Gregorio
Robles +



Michel R.V.
Chaudron *

(*) Gothenburg and Chalmers University

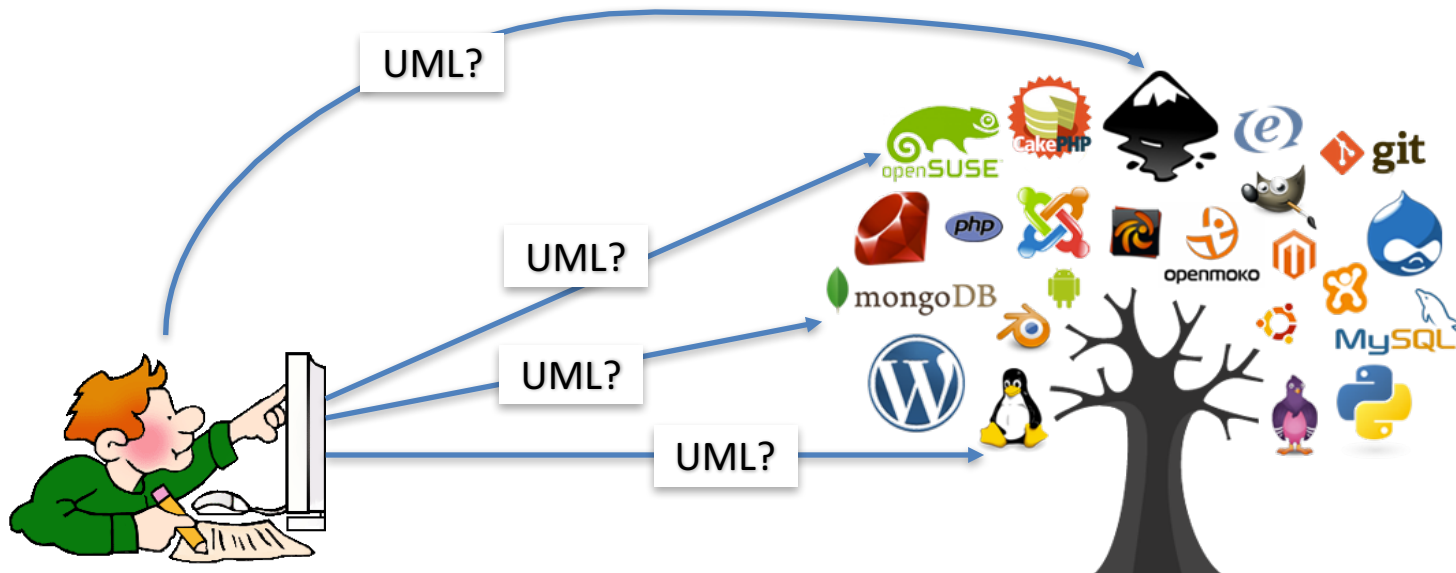
(+) Universidad Rey Juan Carlos

Contents

- Introduction
- Methodology
- Findings
- Discussions

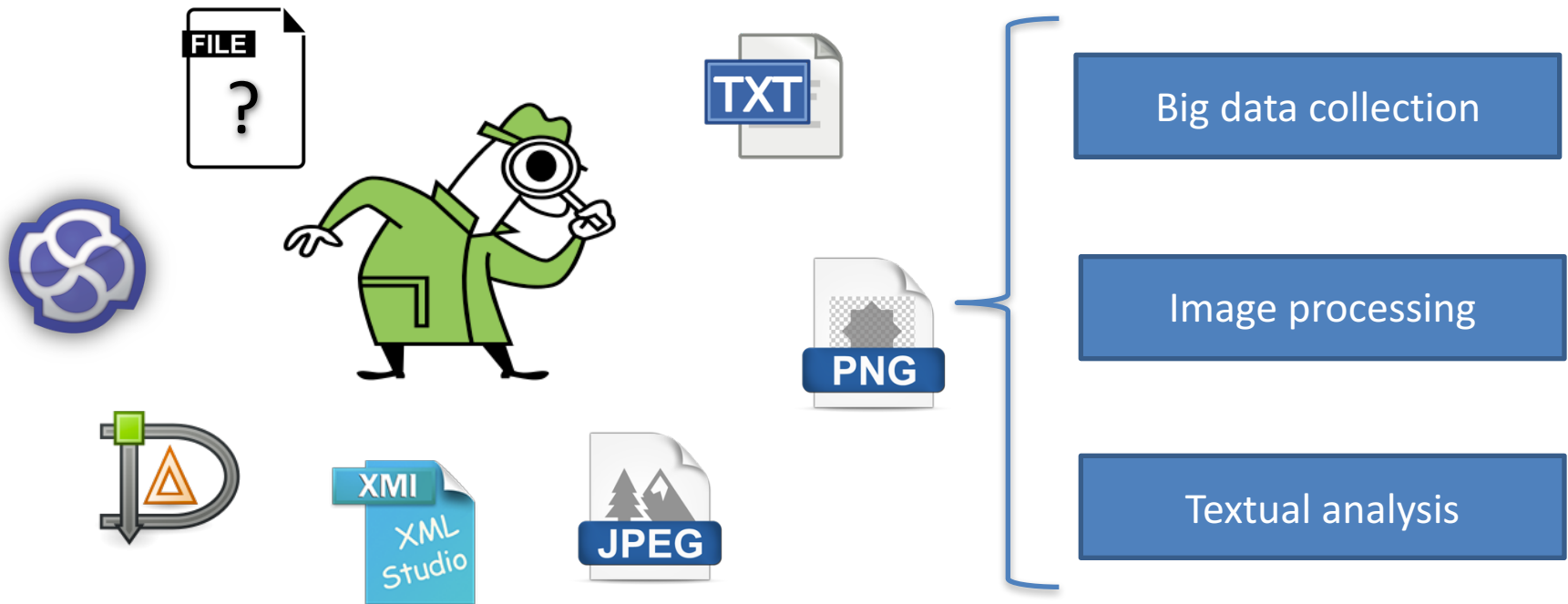
Introduction

- We would like to perform empirical studies of project that use UML
- Little is known about the use of UML in Open Source
- There is no systematic approach to identify UML in OSS



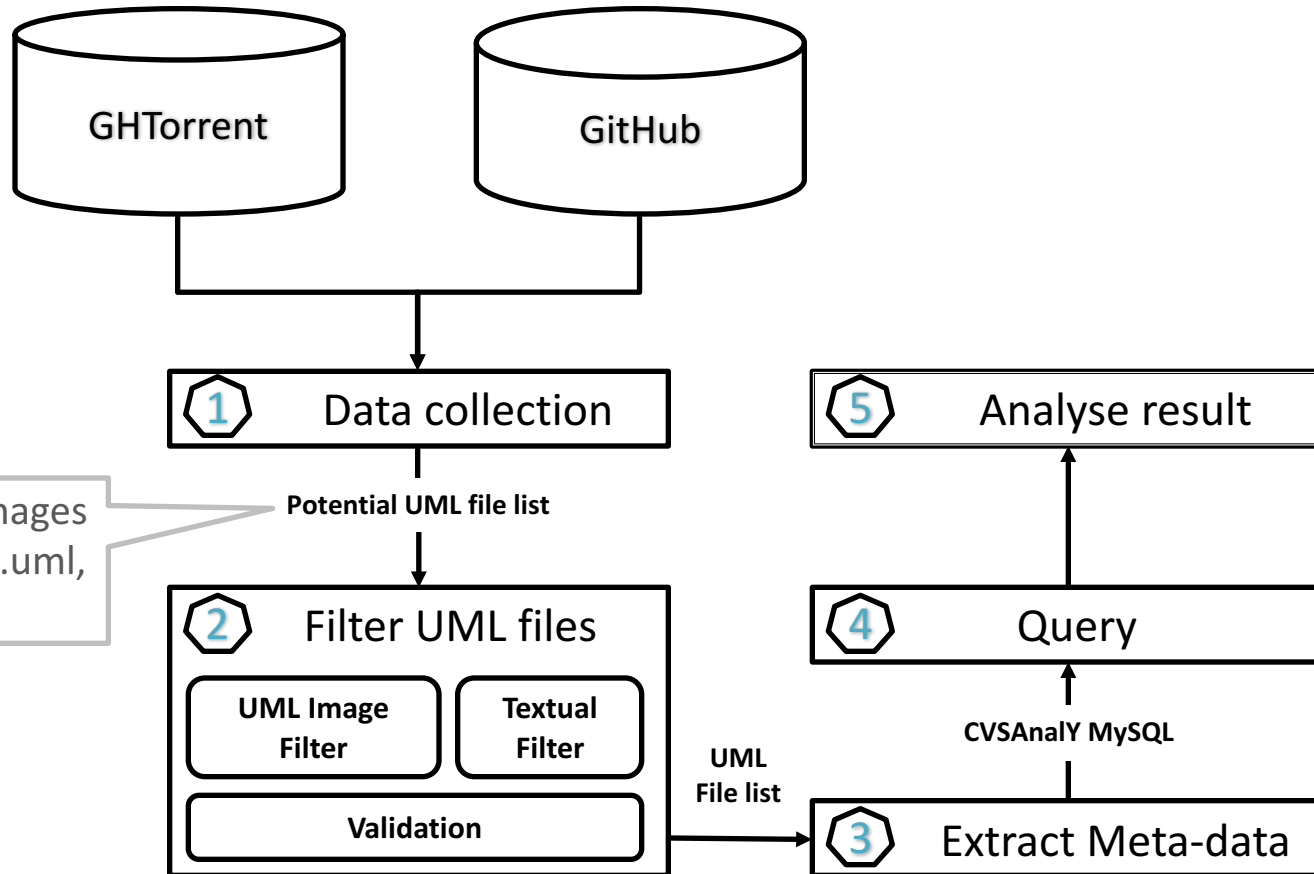
Why is finding of UML diagrams difficult?

- Large variety of projects
- Lack of versioning mechanisms for UML
- Lack of default UML tools and formats



? How to systematically identify open source repositories that contain UML diagrams? – **let's start with GitHub**

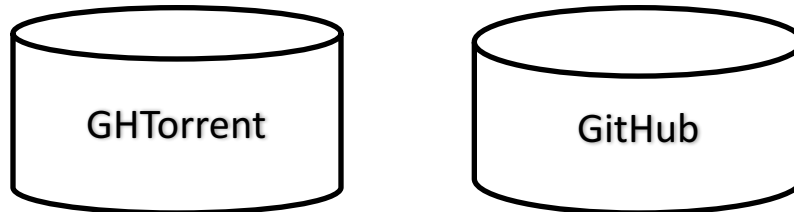
Methodology



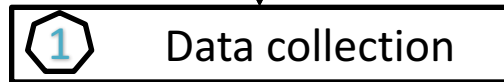
- 1.G. Gousios and D. Spinellis. Ghtorrent: Github's data from a firehose. In Mining Software Repositories (MSR), 2012
- 2.G. Robles et.al. Tools for the study of the usual data sources found in libre software projects. International Journal of Open Source Software and Processes, 1(1):24–45, 2009.
- 3.T. Ho-Quang et.al. Automatic classification of uml class diagrams from images. APSEC '14, pages 399–406, Washington, DC, USA, 2014. IEEE Computer Society.

Some statistics

July 2015

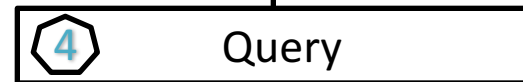
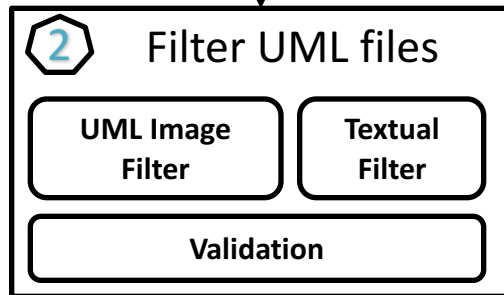


1 240 000 non-forked repos
(10% of GitHub)

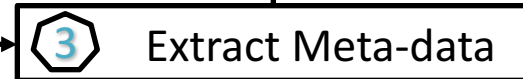


100 702 URLs

Potential UML file list



CVSAnalY MySQL



December 2015

Identified 21 316 UML files
3 295 repositories



Are there GitHub projects that use UML?

Distribution of projects by number of UML files

No. UML files	1	[2-9]	[10-99]	[100,∞)
No. repos	1 947	1 169	158	21

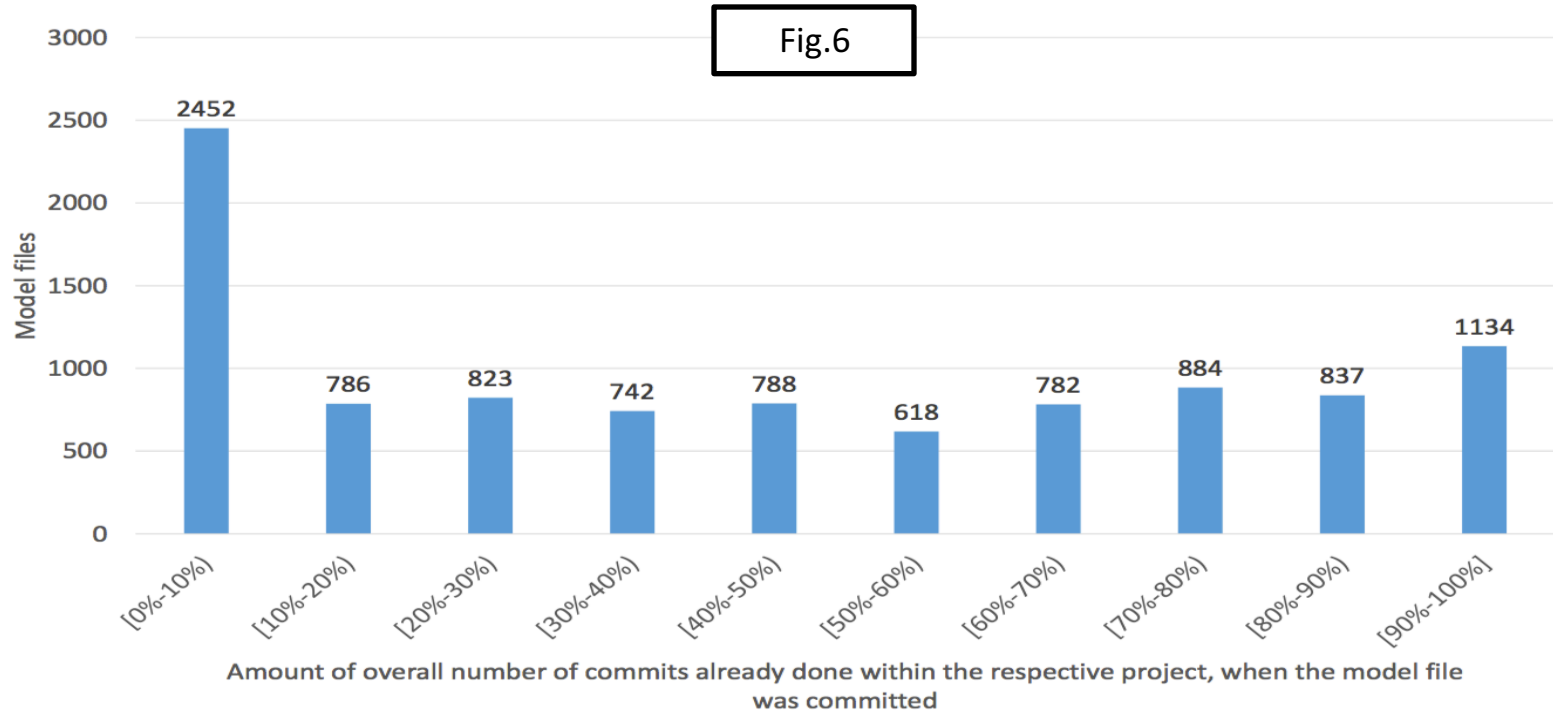
Distribution of UML files by file type

File type	xmi	uml	jpeg	png	gif	svg	bmp
Share	3.4%	44.9%	4.7%	29.6%	16.6%	0.6%	0.2%

We identified 3 295 GitHub repositories
that contain 21 316 UML models.



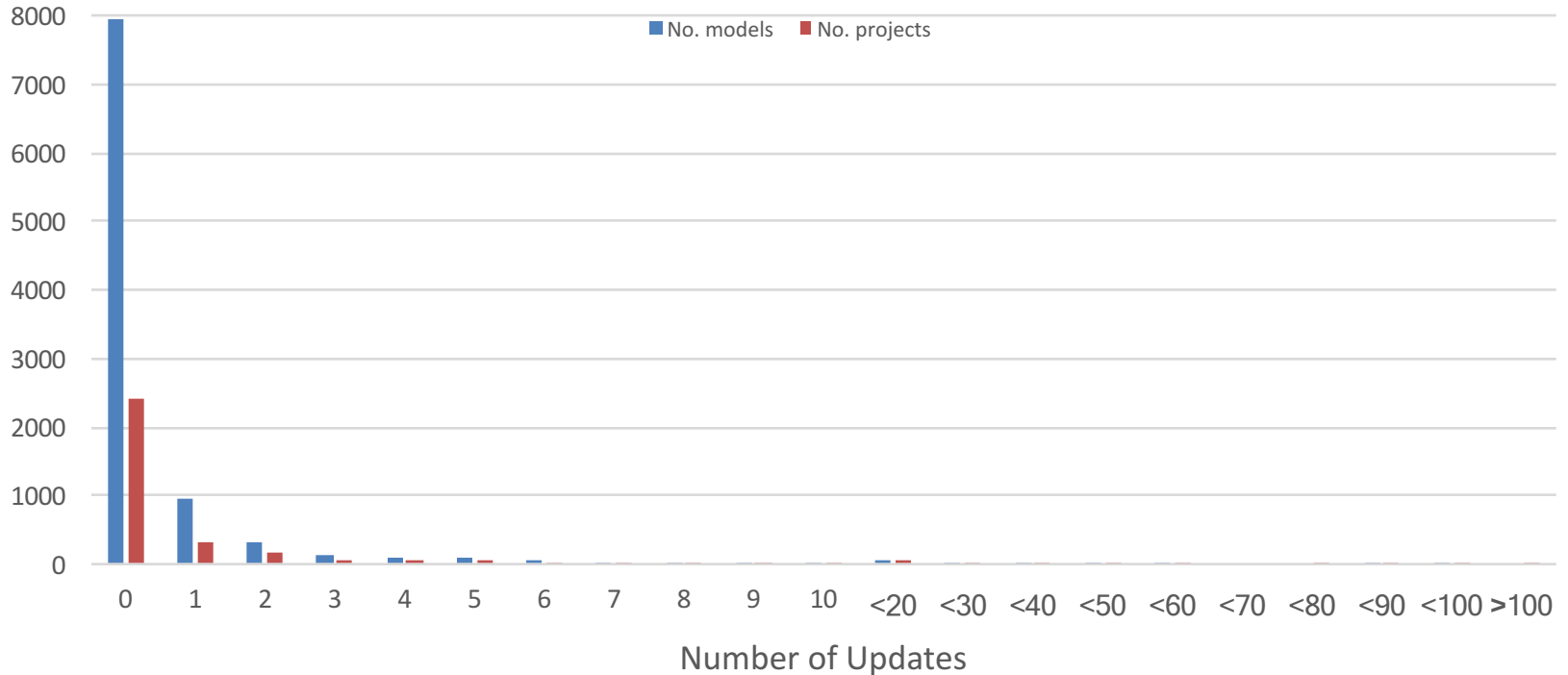
When in the project are new UML models introduced?



UML models are introduced in all active phases of a project with a tendency towards the early phases.



Are there GitHub projects in which the UML models are also updated?



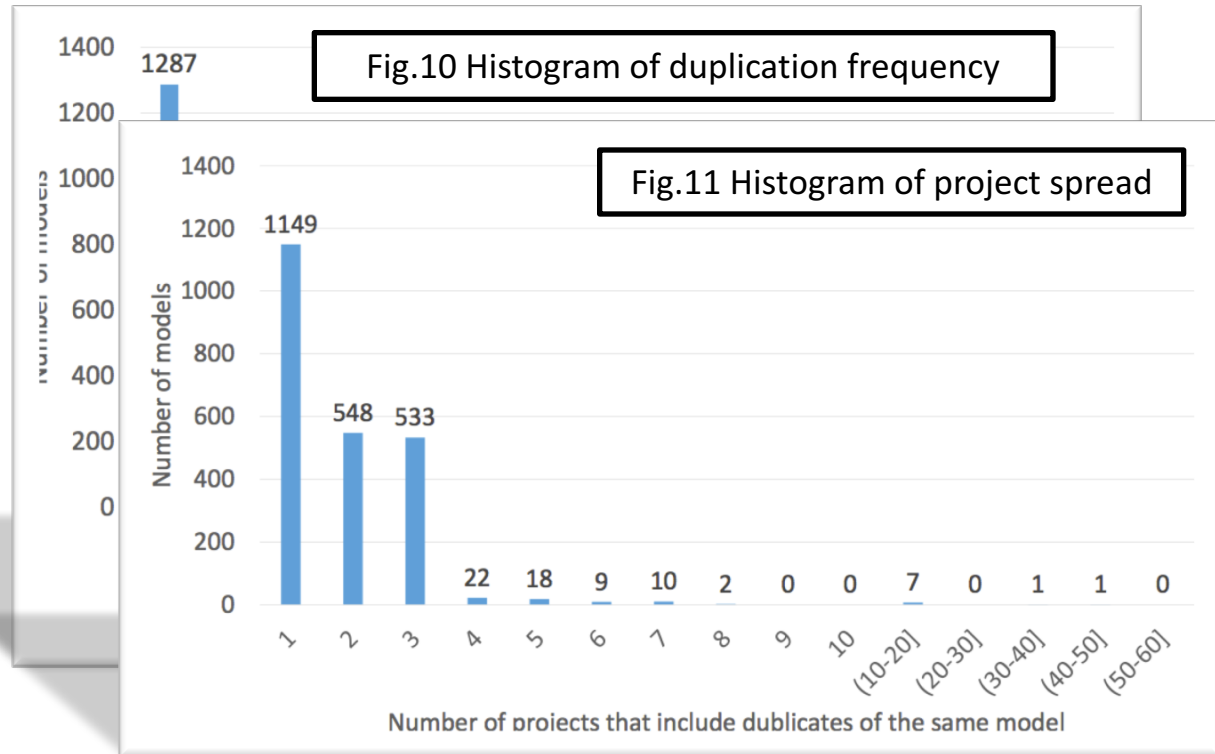
The majority of the UML files (73.33%) are never updated.
26% of the investigated projects updated their UML files at least once.



Are UML files originals?

Out of 21 316 found models

- 18 876 are distinct
- ~ 2300 are duplicates





While most models seem to be unique, a large number of identified distinct models (12%) occur several times. Half of the model duplicates remain within the same project.

Limitations

- Reasons behind model usage?
- Variety in type of projects
e.g. student-projects?

Future research

- Extend study for complete GitHub 
- Find additional model formats
(Powerpoint, PDF, Word) 
- Use the data for investigations:
 - Can UML help to integrate new developers?
 - What files are changed together with changes in architectural models?
 - Does UML contribute to success of open source projects?

Take Away

Offered online (10% of GitHub projects):

- List of 3 295 GitHub repositories which contain 21 000+ UML diagrams
- Replication package of the paper

We believe this dataset will enable many empirical MODELS-studies

Coming soon – end of October
(100% of GitHub projects):

- 24 000+ repositories, 93 000+ UML diagrams (including 35 000+ class diagrams)

For sure you have your own ideas/proposals?!

- What “meta-data” shall we provide about the projects?

Are you looking for models?

lected

35 000+ class diagrams

- open source projects at GitHub

ams can be traced back to the projects, hence it is possible to find project data such as source code, commit messages, commit-dates, more.

d love to hear from you:

you like to use the dataset in your research?

research questions do you recommend us to look at?

ang, T. & Robles, G. & Fernandez, M.A. & Chaudron, M.R.V. (2016). The Quest for Open Source Projects that Use GitHub In proceedings, ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems, Saint-October 2-7, 2016.

Dataset



<http://oss.models-db.com/>

Research Group



Regina Hebig



Trung Ho-Quang



Miguel-Angel Fernandez



Gregorio Robles



Michel R.V. Chaudron