Online Img2UML Repository: An Online Repository for UML Models

Bilal Karasneh[1] and Michel R. V. Chaudron[2,1]

[1] Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands
bkarasne@liacs.nl
[2] Joint Department of Computer Science and Engineering,
Chalmers University of Technology and Gothenburg University, Sweden
chaudron@chalmers.se

Abstract. The Img2UML repository is a repository of UML models. A huge amount of UML models is available on the Internet – mostly in the form of images. This repository aims to offer these UML class diagram as a searchable XMI Database. The information that is in the XMI files is stored in the repository database. This repository will be useful for research as the first corpus of UML models. This repository will provide a good place for researchers and students to study and analyze UML models. This improves UML studying in research, education, and industry. In this paper, we outline the Img2UML repository, illustrate some of the research made possible, and discuss future plans.

Keywords: Model Repository, UML, XMI

1 Introduction

The UML (Unified Modeling Language) language enables the graphical, high-level representation of software. UML models are created during different stages of the software development process. Often, a UML design is the blueprint of the software, and a good UML design helps to realize good software implementations.

Many aspects of software development across the software lifecycle can be measured with a high degree of automation and efficiency. Most software measurements are focused on code instead of on the design of the software, because: 1) Numerous metrics like complexity, maintainability, and readability have been developed for code, but for software design metrics still suffer of many problems. 2) Sharing of the source code artifacts is well supported through platforms such as GitHub [19], and there is no adequate support for sharing modeling artifacts.

One of the main problems of studying UML models is the lack of sharable software development software [1]. In Software Engineering there is a need to share modeling artifacts [2]. Until now, there is no public repository for models. The collection of models from commercial software development is difficult because for different reasons companies like to keep their system design confidential. In open source software, development use of UML is not as common as the (inevitable) use of source code. This makes collecting UML models more difficult, and this difficulty makes empirical research of UML challenging. Moreover, there is no open technology for creating model-repositories as there exist for source code. Many free code repositories are available, which improves the ability of developing code metrics, and facilitates empirical research for source code domain in general.

To facilitate the studying UML models, a set of UML models must be collected. It is challenging to collect UML models because there are a large variety of representations (both graphically and in terms of XMI) of UML models by different UML-CASE (Computer Aided Software Engineering) tools.

In our proposed repository, we start with focusing on one type of UML model, which it is UML class diagram. This selection is done based on the importance of this diagram in software development and its availability. Class diagrams are ubiquitous in UML modeling. UML class diagrams are the most important structural model of the UML, as it shows the static description of the system in terms of classes, relationships and constrains in the relationships [3].

We found that UML models are available in abundance on the Internet, but rather than in CASE-tool format, they are stored in image formats. The problem with image formats is that the model-content of the images cannot be easily extracted out of them. Although many CASE tools support features like creating, modifying and exporting UML models into different formats, current CASE tools cannot recognize UML in images. This inability of CASE tools limits the usability of the availability of UML models in images. For our repository, we are collecting UML class diagram in images from the Internet, and use an image recognition tool [4] that coverts UML class diagram in images into UML models. After this transformation, the tool then saves the images, XMI files and the content of XMI files into the repository. In this way we unlock a huge number of UML class diagrams, which gives a great opportunity for empirical UML research.

The paper is structured as follows: Section 2 describes related work. Section 3 motivates the usefulness of the repository. Section 4 describes the construction of the repository. The conclusion and future work are in Section 5.

2 Related work

Nowadays, a few UML repositories are available. These repositories are supported by CASE tools vendors [17][18] on a commercial basis. Because of the associated costs, these models are not considered attractive from the viewpoint of academic research.

Another kind of repository is a general model repository. In [1], authors proposed repository for model-driven development (ReMoDD) that contains many documented case studies. This repository is a great asset for researchers where they can find many examples of models as well as research studies. However, UML models in ReMoDD are stored as files, so that models are not searchable, and to see a model you have to download it and then open it using compatible CASE tool. In addition, some of case studies do not contain UML models. To complement this, we propose the idea of creating a repository for UML class diagrams as first step towards creating a repository for UML models.

Companies have a huge amount of information at their disposal that is stored in as paper or poorly structured format as PDF, and they need to convert at least most important information into richer format that can be easily searched and modified [5]. In software engineering, this challenge is bigger as software documentation is rich in graphical content. UML models are one of these contents that are mostly available as images in software documentation and on the Internet. The problem is the lack of mapping from a pixel-based diagram to the underlying engineering model conveyed by the diagram [6].

The area of converting engineering diagrams into engineering models has received some attention [5-11]. Some of this research is oriented at recognizing graphic objects or symbols in images [7-10]. Other research aims at recognizing entire models in images [11]. For specifically for UML diagrams, researchers have focused on converting hand-drawn sketching of UML class diagram into models [12-15]. In general, hand-drawn tools are an easy and fast way to create and (re)draw UML class diagram than UML CASE tool. However, redrawing UML models from paper in order to enable editing them again is very time-consuming. The algorithms that are used in recognizing UML diagrams in hand-drawn tools typically make use of information regarding the movement and order of drawing elements in the diagram. This makes these algorithms unsuitable for extracting UML models from 'finished' diagrams.

In our earlier work [4] we proposed the Img2UML tool that converts UML class diagrams (including class names, attributes, relationships) that are represented in image format into XMI (version 1.1, the UML version is 1.3). The resulting XMI files generated by the tool are compatible with StarUML [16]. This tool also contains functionality to save models as XMI files.

3 Usefulness of the repository

This new Img2UML repository aims to be a source for empirical studies of UML class diagrams. This repository opens up a lot of uses:

- It can be the basis for corpus studies for UML modeling, all available models are validated manually, and any mistake in the recognition can corrected manually.
- It can be the basis of metrics used in benchmarking for quality assurance of UML models, such as the average number of classes per model and the average number of attributes and operations per class. This provides an empirical basis for UML quality assurance.
- It can serve as a source of UML models that can be used in and shared across empirical studies in UML modeling
- It can serve as a source of examples of UML design that can be used for educational purposes, e.g. learning UML by examples.

This repository already contains 1000 class models. The repository can be used to analyze class diagrams, measure qualities, study typical flaws and their frequency of occurrence, compare quality-models, etc. Although more rare, the availability of different versions of models for one software system provides an opportunity to study the evolution of versions of the class designs. The repository can also be used for studying UML class diagram in software engineering classes. Students can reuse available class diagrams, share their knowledge, and engage in discussions about models.

Our system offers functionality for querying and searching the repository of models based on different keys such as model information (class name, attributes, etc.). Models in the repository can be classified and analyzed automatically by using some queries on the repository. For example, which class diagram contains a high number of classes, a high number of relationships (dependency or inheritance), or some design pattern. These queries can show common characteristics of class diagrams.

The content of ReMoDD and our repository are different in: First, ReMoDD contains documents, model files and codes, and in our repository only models are available. Second, all models in our repository are editable and searchable. Third, our models are collected from software documents on the Internet. However, models in ReMoDD seem to come from (industrial) case studies. Forth, our repository supports querying models, because models information (contents of XMI) such as names of classes, attributes and operations with relationships are stored in the database. Fifth, although our repository contains models that are created using different CASE tools, it is not obligatory to have these CASE tools to use these models because all XMI files are compatible with StarUML. Moreover, we could easily add a feature for exporting to other versions of XMI.

4 Repository Description

4.1 Collecting UML class diagrams

Different UML class diagram images are collected from the Internet using Google Image Search. These images vary in color, type, size, and resolution. Images are collected together with their URLs. These URL’s are used as keys in the database and are used to prevent including duplicates of images. After this, additional manual checking is performed to assure that indeed no duplicate images end up in the repository.

4.2 Inserting model information into the database

The process of inserting models in the database can be divided into two parts: First, the UML class diagram is extracted from an image. For this, the Img2UML tool converts class diagrams in images into XMI files. Second, the XMI content is saved into the database. From XMI, classe names, attributes names, operations names and relationship types are read, and saved in the database. Figure 1 shows the structure of the database. The repository contains 10 tables, where the image_Table contains image_IDs, the available UML class diagram images, URLs of images and images properties such as width, height, and resolution. The xmi_Table contains XMI files and generals comments about the models. This comments will improved to be more classified, as comments about layout, understandability, complexity, recognition, etc.

Both attributes_Table and operations_Table contains attributes names and operations names and where this attributes and operations are available in which classes. The remains tables are related to the relationships, where each relationship is saved in the in details, for example in the generalization_Table, the generalization_Child and generalization_Parent show inheritance relationships between classes and xmi_ID shows that this relation is available in which model.

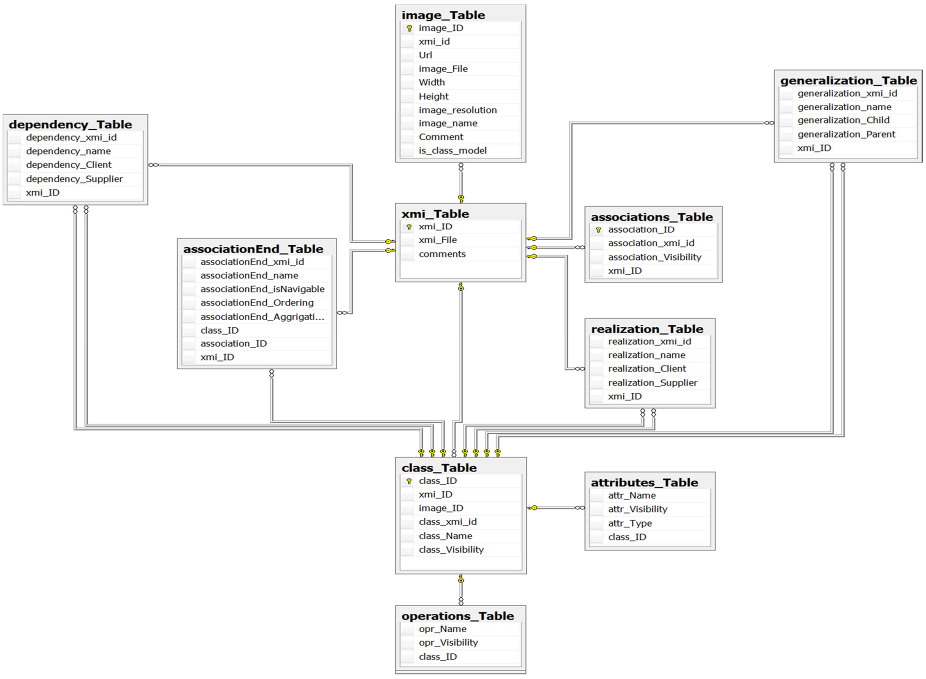

Fig. 1. Img2UML database structure

5 Conclusion and Future Work

In this paper, we proposed the Img2UML repository, a repository for UML diagrams based on UML class diagram in images that have been converted to XMI. The repository contains UML class diagrams images that are collected from the internet. The repository also contains the images' URLs and the corresponding XMI files that are generated via special tool created for recognizing UML models from diagrams. A web-based user interface will make the repository more available and accessible. The goal of the repository is to be a basis for UML models that can be used and shared across empirical studies.

For future work, we will evaluation different aspects of the repository. Also we aim to develop an API for uploading more UML class diagrams in images by users. More information and classification about available UML class diagrams will be supported, like information about the related software development project.

6 References

1. France, R., Bieman, J., Cheng, B.H.: Repository for model driven development (ReMoDD). Models in Software Engineering, pp. 311-317. Springer (2007)
2. Buse, R.P., Zimmermann, T.: Information needs for software development analytics. In: Proceedings of the 2012 International Conference on Software Engineering, pp. 987-996. IEEE Press, (2012)
3. Maraee, A., Balaban, M.: Efficient recognition of finite satisfiability in UML class diagrams: Strengthening by propagation of disjoint constraints. In: Model-Based Systems Engineering, 2009. MBSE'09. International Conference on, pp. 1-8. IEEE, (2009)
4. Karasneh, B., Chaudron, M. R. V.: Extracting UML Models from Images. In: 5th International Conference on Computer Science and Information Technology, CSIT 2013. (2013)
5. Tombre, K., Lamiroy, B.: Graphics recognition-from re-engineering to retrieval. In: Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, pp. 148-155. IEEE, (2003)
6. Fu, L., Kara, L.B.: From engineering diagrams to engineering models: Visual recognition and applications. Comput. Aided Des. 43, 278-292 (2011)
7. Barrat, S., Tabbone, S.: A Bayesian network for combining descriptors: application to symbol recognition. International Journal on Document Analysis and Recognition (IJDAR) 13, 65-75 (2010)
8. Barrat, S., Tabbone, S., Nourrissier, P.: A bayesian classifier for symbol recognition. In: Seventh International Workshop on Graphics Recognition-GREC'2007. (2007)
9. Luqman, M.M., Brouard, T., Ramel, J.-Y.: Graphic symbol recognition using graph based signature and bayesian network classifier. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, pp. 1325-1329. IEEE, (2009)
10. Yang, S.: Symbol Recognition via Statistical Integration of Pixel-Level Constraint Histograms: A New Descriptor. IEEE Trans. Pattern Anal. Mach. Intell. 27, 278-281 (2005)
11. Yu, Y., Samal, A., Seth, S.C.: A System for Recognizing a Large Class of Engineering Drawings. IEEE Trans. Pattern Anal. Mach. Intell. 19, 868-890 (1997)
12. Chen, Q., Grundy, J., Hosking, J.: SUMLOW: early design-stage sketching of UML diagrams on an E-whiteboard. Softw. Pract. Exper. 38, 961-994 (2008)
13. Hammond, T., Davis, R.: Tahuti: a geometrical sketch recognition system for UML class diagrams. ACM SIGGRAPH 2006 Courses, pp. 25. ACM, Boston, Massachusetts (2006)
14. Lank, E., Thorley, J., Chen, S., Blostein, D.: On-line Recognition of UML Diagrams. In: Proc. 6th ICDAR (2001) 356-360
15. Lank, E., Thorley, J.S., Chen, S.J.-S.: An interactive system for recognizing hand drawn UML diagrams. Proceedings for CASCON 2000; 2000. p. 7
16. StarUML - http://staruml.sourceforge.net/en/
17. Enterprise Architect - http://www.sparxsystems.com/
18. Visual Paradigm - http://www.visual-paradigm.com/
19. GitHub - https://github.com/