1 INTRODUCTION

Software engineering has been a conventional methodology that is followed for the development of software. It is still not well adopted by the scientific community so far. On top of that the topic of secure software is emerging without being handled appropriately.

Driven by the reproducible science paradigm and ever growing research networks, scientific software is made available to a broader range of users. Despite an alarming amount of vulnerabilities has been made public over the last years, software is still shared without being hardened with respect to security issues. Thus scientific software might introduce attack vectors to target research institutes.

Our goal is to come up with a collection of guidelines and tools to apply during development to create secure software without in-depth knowledge of software security.

At DLR we have access to a range of software repositories\(^1\) as well as many ongoing research projects where software is developed.

In this paper we want to outline our research as follows:

- We characterize software development at DLR to illustrate the context of our research (Sect. 2).
- We describe the need to address the security issues in DLR with better insight into secure software engineering based on data analysis (Sect. 3).
- We present our strategies to analyze software development processes and to create a catalog with tools and guidelines that supports secure software development (Sect. 4).

2 SOFTWARE DEVELOPMENT AT DLR

The German Aerospace Center (DLR) is one of the largest research organizations in Germany. Over 8000 scientists are researching in the fields of aeronautics, space, transportation and energy. In these fields many tasks rely on computer systems. This involves individual software developed by domain experts in multiple programming languages across different platforms.

More than 2000 people at DLR are occupied with software development in part-time or full-time [1]. Most of the developers have no training in software development, only very few have deeper knowledge about systematic development of sustainable software, the means of software engineering, or secure software. Typical

\(^1\)This includes but is not limited to version control systems as well as issue trackers and continuous integration systems.
development team sizes range from one up to 20 persons, in average being one scientist supported by interns and perhaps a Ph.D. student.

The combination of being domain scientists and small team sizes makes the amount of needed knowledge about software engineering and security a disproportional overhead for a project.

3 SOFTWARE SECURITY IN DLR

As DLR is well known for its expertise we have lots of cooperating software. As data we use the software projects of DLR development. To accomplish this goal we apply methods from data science to capture provenance data, the variety of used IDEs at DLR, logger approach. While IDE extensions are our preferred way of automated security auditing, however they are mostly very time consuming and produce rather vague results in contrast to the static analysis approach. The static analysis can be evaluated based on manual audits, syntax tree analysis, and intermediate language analysis. We will focus on the latter-most analysis approach as the existing vulnerabilities from databases like CVE\(^2\) or the exploitation frameworks can be transferred into intermediate language. These then can be used as examples for vulnerable or exploitable code.

Unexperienced developers at DLR have seen the deployment of software in the cloud as a solution to decouple the execution environment and security a disproportional overhead for a project.

Hence in our first step we rely on the documented process and incorporate easy to adapt techniques like surveys and repository mining to augment this data.

(2) We monitor the quality of the software using manual and automated audits. Therefore we investigate and improve existing static and dynamic security analysis methods. The dynamic analysis provides the most universal and versatile way of automated security auditing, however they are mostly very time consuming and produce rather vague results in contrast to the static analysis approach. The static analysis can be evaluated based on manual audits, syntax tree analysis, and intermediate language analysis. We will focus on the latter-most analysis approach as the existing vulnerabilities from databases like CVE\(^2\) or the exploitation frameworks can be transferred into intermediate language. These then can be used as examples for vulnerable or exploitable code.

(3) We conduct experiments to identify process properties that have an effect on the security of the resulting software. Factors such as security-focused requirements engineering or a special security testing phase promise a high impact. However we also want to experiment with other approaches such as threat modeling and special trainings for developers. To allow comparison of software quality across projects, we introduce a software security scoring system based on automated software analysis.

5 CONCLUSION

In order to improve software development processes we started a new research group. Our aim is to optimize process properties using approaches from data science. We include two main sources of data: the provenance of software processes and a score for the software security of that artifact.

We presented some strategies for collecting both of them:

- We plan to use repository mining as a source for process information. This should also help to identify missing information that needs to be recorded with another approach
- To augment the mined data, we plan to introduce developer surveys. In a later step we also plan to implement process recording extensions for IDEs.
- We plan to derive a common security scoring system based on existing dynamic and static analysis techniques.
- We develop a new data driven static analysis approach based on the intermediate language representation of source code.

We will apply these approaches in real projects in the environment of DLR, which gives us large datasets that can be used to improve results.

REFERENCES


\(^2\)https://cve.mitre.org/