A Dataset of Software Bill of Materials for Evaluating SBOM Consumption Tools

Rio Kishimoto[1] Tetsuya Kanda[2] Yuki Manabe[3] Katsuro Inoue[4] Shi Qiu[5] Yoshiki Higo[1]
[1]The University of Osaka, Japan [2]Notre Dame Seishin University, Japan [3]The University of Fukuchiyama, Japan
[4]Nanzan University, Japan [5]Toshiba Corporation, Japan

Software Engineering Laboratory

SBOM & SPDX

Software management using SBOM is recommended

SBOM (Software Bill of Materials)
List of components that make up software, including information on dependent libraries (name, version, dependency etc.)

SPDX
One of the major formats of SBOM
Developed by Linux Foundation
Standardized as ISO/IEC 5962:2021

Tools related to SBOM

SBOM generation tools
- Generate an SBOM from software information
- These tools have been studied extensively[1, 2]

SBOM consumption tools
- Support software management using SBOMs
  - Detect vulnerability / license violation
  - View/Edit SBOMs
- **Research on SBOM consumption tools remains limited**

Lack of Datasets of SBOMs

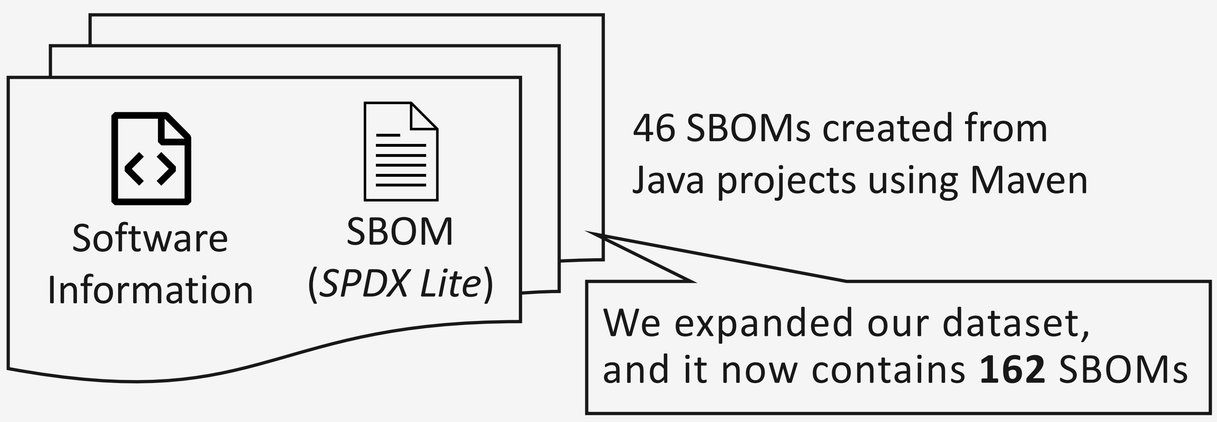
- No publicly available dataset to evaluate SBOM consumption tools
- Examples of SBOM typically describe software with few dependencies and their utility for evaluating tools is limited

To effectively evaluate SBOM consumption tools, accurate and well-structured SBOMs is necessary

To address this gap, we present a dataset of SBOMs

Dataset Overview

- Pairs of software information (name, URL etc.) and its SBOM
- Created from Java projects on GitHub
- SBOMs are *SPDX Lite* compliant
  - *SPDX Lite*: a set of mandatory fields based on actual workflows in industries
  - By following this profile, we ensure that **the SBOMs contain the critical information needed for the major use cases** (vulnerability assessment and license management)

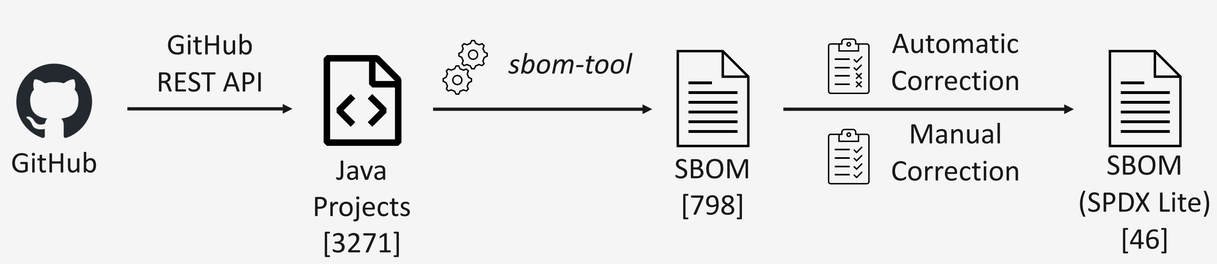


15(S)M-Tu

Preprint
https://arxiv.org/abs/2504.06880

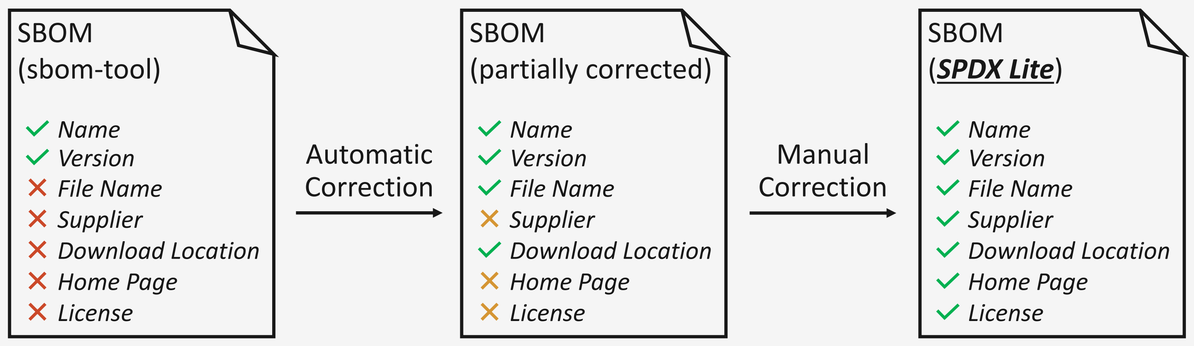
SBOM Creation Steps

1.Collect Java projects from GitHub
2.Generate SBOM using an SBOM generation tool
3.Correct automatically
4.Correct manually

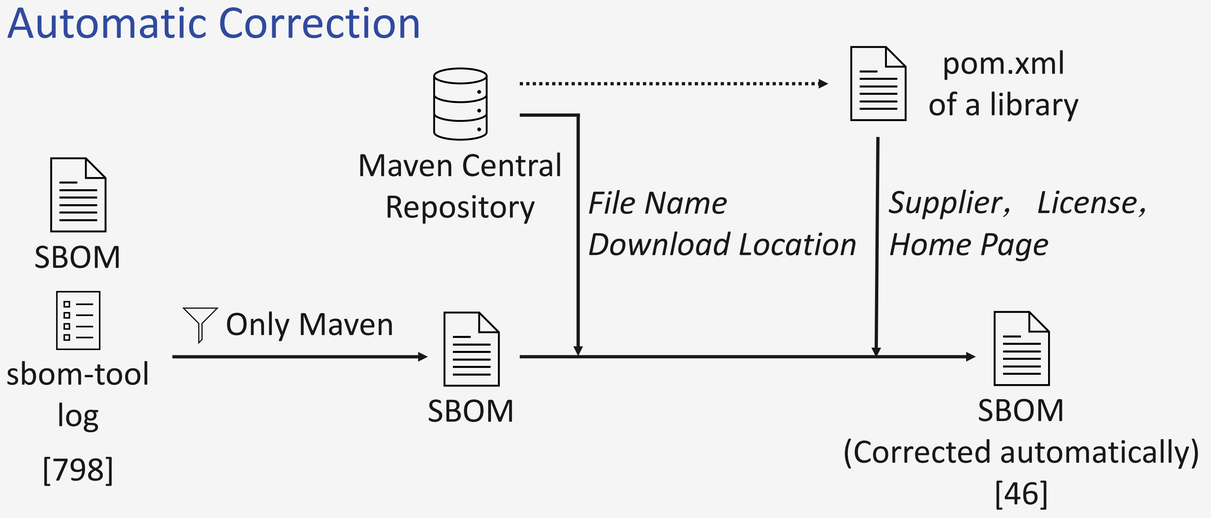


SBOMs produced by *sbom-tool* do not contain some pieces of information

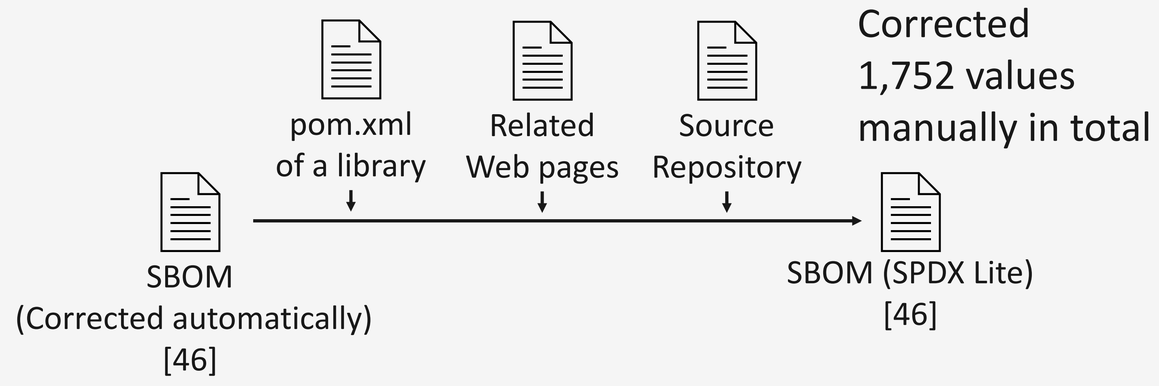
Add missing dependent library information to the SBOMs and make them *SPDX Lite* compliant



Automatic Correction



Manual Correction



Conclusion

Present a Dataset of SBOMs
- 46 SBOMs created from Java projects on GitHub
- SBOMs are compliant with SPDX Lite profile
- Provide quality-assured SBOM samples and facilitate evaluation of SBOM consumption tools

Future Work
- Expand the dataset to include a broader range of projects across various programming languages

References:
[1] On the Way to SBOMs: Investigating Design Issues and Solutions in Practice (Bi et al, 2024, TOSEM)
[2] Challenges of Producing Software Bill of Materials for Java (Balliu et al, 2023, IEEE Security & Privacy)