Bachelorarbeit von Yan Wittmann

an der Hochschule Mannheim, Fakultät für Informatik, im SS2025
in Kooperation mit der metaeffekt GmbH.

Hintergrund

Die metaeffekt GmbH entwickelt eine Software zur automatisierten Identifikation von Schwachstellen in Softwareprodukten, die frei unter einer offenen Lizenz verfügbar ist. Ein langfristiges Ziel ist die vollständige Automatisierung der Schwachstellenidentifikation.

Für die Schwachstellenzuordnung werden öffentlich zugängliche Datenquellen wie die NVD, OSV-Datenbanken und Security Advisories unterschiedlicher Autoritäten, Institutionen und Herstellern verwendet. Um eine automatisierte Verarbeitung dieser Dokumente zu ermöglichen, geben diese ihre Produktdaten mit verschiedenen Standards wie CPE, PURLs oder sogar proprietäre IDs an. Dies macht individuelle Anpassungen am Code für jede neue Datenquelle erforderlich.

Auch die Eingabedaten in Form eines Software-Inventars, die aus der Software-Kompositionsanalyse stammen, variieren oft, da unterschiedliche Betriebssysteme, Paketmanager oder Projektkonventionen zu uneinheitlichen Darstellungen derselben Software führen.

Derzeit analysiert der Schwachstellen-Scanner diese unterschiedlichen Produktidentifikatoren mithilfe individueller Prüfregeln. Der Erfolg dieses Ansatzes ist jedoch stark formatabhängig und erreicht bei den meisten nur begrenzte Erfolgsraten bei der automatischen Zuordnung. Einige Formate wie PURLs lassen sich leichter verarbeiten, während CPEs oder Microsoft Produkt-IDs oft nur ungenau oder gar nicht zugeordnet werden können.

Um die Ergebnisse des automatisierten Prozesses händisch ergänzbar zu machen, wurde vor einigen Jahren ein manueller Prozess namens “Produkt-Korrelation” eingeführt, der durch ein Korrelations-Team gepflegt wird. Hierbei werden manuelle Korrekturen in YAML-Dateien dokumentiert, die durch ein einfaches Web-UI unterstützt werden. Dieses System stößt jedoch zunehmend auf Skalierungsprobleme, da die YAML-Dateien mit teils mehreren tausend Zeilen unübersichtlich und schwer wartbar geworden sind. Es besteht Bedarf für ein zukunftsfähiges, verbessertes Format und mit einem neuen Konzept für die Modellierung von Produkten und deren Beziehungen.

Ziel der Bachelorarbeit

Auf Basis der dargestellten Herausforderungen im Hintergrund ergibt sich folgende Zielsetzung für meine Bachelorarbeit in Kooperation mit der metaeffekt:

- Analyse des Ökosystems der Produktidentifikationsstandards.
- Analyse weiterer öffentlicher Tools im selben Kontext hinsichtlich deren Methoden zur Datenverarbeitung und -nutzung.
- Vergleich der internen Eingabedaten mit externen Identifikationsformaten.
- Untersuchung der Schwächen des bestehenden YAML-Korrelationsformats.
- Konzeption und Implementierung eines skalierbaren und zukunftssicheren Korrelationsformats.
- Anwendung der gewonnenen Erkenntnisse, um die automatisierte Produktidentifikation zu verbessern und weiterzuentwickeln.

Literatur und Quellen

Im Rahmen dieser Arbeit wurden unter anderem bereits diese relevanten Quellen untersucht:

- Software Identification Ecosystem Option Analysis
- Graph-Based CPE Matching for Identification of Vulnerable Asset Configurations
- Vortrag: Universal Software Product Identity (Thomas Schmidt, FIRSTCON23)
- DependencyCheck (GitHub-Projekt)

Spezifikationsdokumente

PURL, CPE 2.2/2.3, OSV, CSAF, etc.

Zitate

Durch ein Gespräch mit Thomas Schmidt, BSI:

> Product Identification presents one of the greatest challenges in the field of vulnerability management. If it is not solved, complete automation of vulnerability matching is not possible.
>
> CSAF 3.0 will not be published until it is solved.

Naming Things is Hard:

> There are only two hard things in Computer Science: cache invalidation and naming things.