netidee

PROJEKTE

CrOSSD

Towards a Critical Open-Source Software Database

Zwischenbericht | Call 17 | Projekt ID 6252

Inhalt

1 Einleitung

Open-Source-Software treibt eine breite Palette von Softwareanwendungen an und bildet die Grundlage für viele Dienste im Internet. Trotz dieser wichtigen Rolle stehen OSS-Projekte oft vor Herausforderungen in Bezug auf Nachhaltigkeit und Wartung. Um diesem Problem zu begegnen, werden in CrOSSD kritische OSS-Projekte identifiziert und bewertet, um so einen umfassenden Überblick über deren „Gesundheit“ zu bieten. Dies geschieht mithilfe verschiedener geeigneter Metriken und automatisierter Analysen. Die verwendeten Metriken umfassen Stabilität, Widerstandsfähigkeit, Sicherheit und Compliance.

Das Projektziel ist der Aufbau einer Plattform, die eine große Anzahl von OSS-Projekten überwacht und bewertet. Dies ermöglicht es verschiedenen Interessengruppen, wie den Projekteigentümer:innen selbst, der gesamten OSS-Community, oder z. B. anderen Entwickler:innen, Unternehmen und Fördergebern, informierte Entscheidungen zu treffen.

In diesem Zwischenbericht präsentieren wir die Forschungsdesign-Methodik und die ersten Implementierungsergebnisse von CrOSSD: Recherche und Definition von Metriken (siehe Arbeitspaket 3 - Metriken: Recherche, Definition und Operationalisierung) sowie die Implementierung von Metriken und erste Analysen (siehe Arbeitspaket 4 – Implementierung und Crawling).

Obwohl es bereits bestehende Forschungsprojekte und Initiativen gibt, die ähnliche Ziele verfolgen und Metriken sowie bewährte Praktiken anbieten (einige sogar mit Bewertungen ähnlich unserer Grundidee), hebt sich CrOSSD davon ab, indem es ein umfassenderes Verständnis für die Gesundheit von Projekten bieten soll. Darüber hinaus bietet keine der existierenden Initiativen kontinuierliche Überwachung und Reporting an. Um unsere Idee zu evaluieren, stellten wir die Projektidee bereits in einem frühen Stadium in der Community zur Diskussion, indem wir unseren derzeitigen Aufbau von CrOSSD auf wissenschaftlichen Konferenzen vorstellten (siehe Arbeitspaket 2 - Projektmanagement, Outreach und Dissemination).

2 Status der Arbeitspakete

2.1 Arbeitspaket 1 - Detailplanung und Formales am Projektstart

Arbeitspaket 1 wurde bereits mit April 2023 erfolgreich abgeschlossen. Die folgenden Arbeiten wurden in diesem initialen Arbeitspaket durchgeführt:

- der Vertrag wurde unterschrieben,
- der Detailprojektplan wurde erstellt und abgenommen,

- eine detaillierte Liste der geplanten Projektergebnisse mit Lizenz und Ort der öffentlichen Bereitstellung wurde erstellt und abgenommen,
- die Projekt-Website gingen Betrieb und ein erster Blogeintrag wurde erstellt, und

die erste Förderrate wurde beantragt.

2.2 Arbeitspaket 2 - Projektmanagement, Outreach und Dissemination

Das Projektmanagement verlief bislang erwartungsgemäß und ohne größere Hindernisse, sodass hier keine Umplanungen oder Ähnliches notwendig waren.

Im Blog unter https://www.netidee.at/crossd haben wir plangemäß bislang drei Beiträge veröffentlicht:

- Im Feber 2023 haben wir unser Projekt zunächst einmal vorgestellt, sowohl bzgl. der Motivation und grundlegenden Idee als auch bzgl. der Systemarchitektur und Konzeption unserer Infrastruktur: https://www.netidee.at/crossd/introducing-crossd
- Im Mai 2023 haben wir unser Projektteam vorgestellt. Zusätzlich zu den bereits bei der Antragstellung eingeplanten drei Projektmitarbeitern (Sebastian Neumaier, Lukas Daniel Klausner und Tobias Dam) sind auch zwei FHStP-Studierende (Matthias Kopeinig und Jacqueline Schmatz) in unser Team hinzugekommen, die ihre Abschlussarbeiten im Rahmen von CrOSSD geschrieben und sich mit Metriken und Analysen befasst haben (mehr dazu im nächsten Abschnitt): https://www.netidee.at/crossd/our-team
- Im August 2023 schließlich haben wir die Arbeit eines der zwei Student:innen vorgestellt. Matthias Kopeinig hat im Rahmen seiner Abschlussarbeit untersucht, welche Erkenntnisse eine Analyse der Firmenzugehörigkeiten von Contributors in einem Sample von 100 GitHub-Projekten offenbaren kann: https://www.netidee.at/crossd/impact-companies-open-source

Weiters haben wir auch erste wissenschaftliche Publikationen veröffentlicht:

- Tobias Dam und Sebastian Neumaier haben sich in „Towards Measuring Vulnerabilities and Exposures in Open-Source Packages“ mit Sicherheitsaspekten im Kontext unseres Projekts befasst. Das Paper wurde Anfang Mai von Sebastian Neumaier auf der 5th International Data Science Conference an der UWK in Krems präsentiert und erscheint demnächst in den Konferenzproceedings (siehe https://idsc.at/proceedings/); ein Preprint ist unter https://arxiv.org/abs/2206.14527 verfügbar.
- Tobias Dam, Lukas Daniel Klausner und Sebastian Neumaier haben weiters im Short Paper „Towards a Critical Open-Source Software Database“ das Projekt vorgestellt, um auch in der wissenschaftlichen Community erstes Networking zu betreiben. Das Paper wurde Anfang Mai von Tobias Dam auf der ACM Web Conference 2023 in Austin, Texas, vorgestellt und ist bereits in den Konferenzproceedings erschienen: https://doi.org/10.1145/3543873.3587336

- Eine dritte Publikation, die den Stand der Forschung zu quantitativen Metriken ausführlich analysieren wird, ist derzeit unter Mitarbeit von Jacqueline Schmatz in Vorbereitung.

Weiters haben wir die Projektwebsite https://crossd.tech/ grundlegend überarbeitet.
Zuletzt wurde der Zwischenbericht fristgerecht abgeschlossen und eingereicht.

2.3 Arbeitspaket 3 - Metriken: Recherche, Definition und Operationalisierung

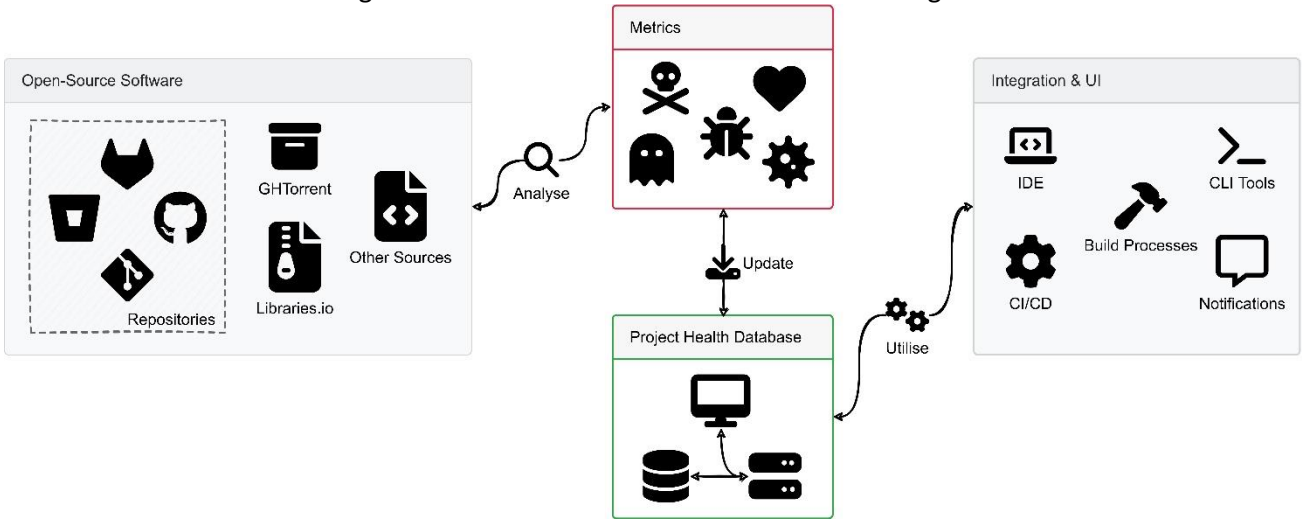
In diesem Arbeitspaket haben wir uns gemäß unserem ursprünglichen Plan zunächst einmal auf quantitative Metriken fokussiert.

In einer systematischen Literaturrecherche haben wir 26 relevante Papers gefunden, die sich mit OSS-Gesundheit oder -Qualität und Metriken oder Kriterien befassen. Wir haben zunächst die unterschiedlichen Gesundheits- und Qualitätsbegriffe sowie die Metriken und Analyseverfahren aus diesen Papers extrahiert und strukturiert aufgeschlüsselt.
Im nächsten Schritt haben wir die verschiedenen Konzeptionen von Gesundheit und Qualität gesichtet, geclustert und codiert, um so eine umfassende Taxonomie zentraler Aspekte von OSS-Gesundheit zu erhalten.
In der Folge haben wir aus den erfassten Metriken jene selektiert, die erstens gewissen Nützlichkeits- und Pragmatismusheuristiken genügen, insbesondere aber die zuvor erfassten Aspekte möglichst gut abdecken. Diese Metriken werden in der Folge nun in unserem System implementiert und erfasst.

Die Resultate dieses Prozesses werden weiters aktuell auch zur wissenschaftlichen Publikation aufbereitet.

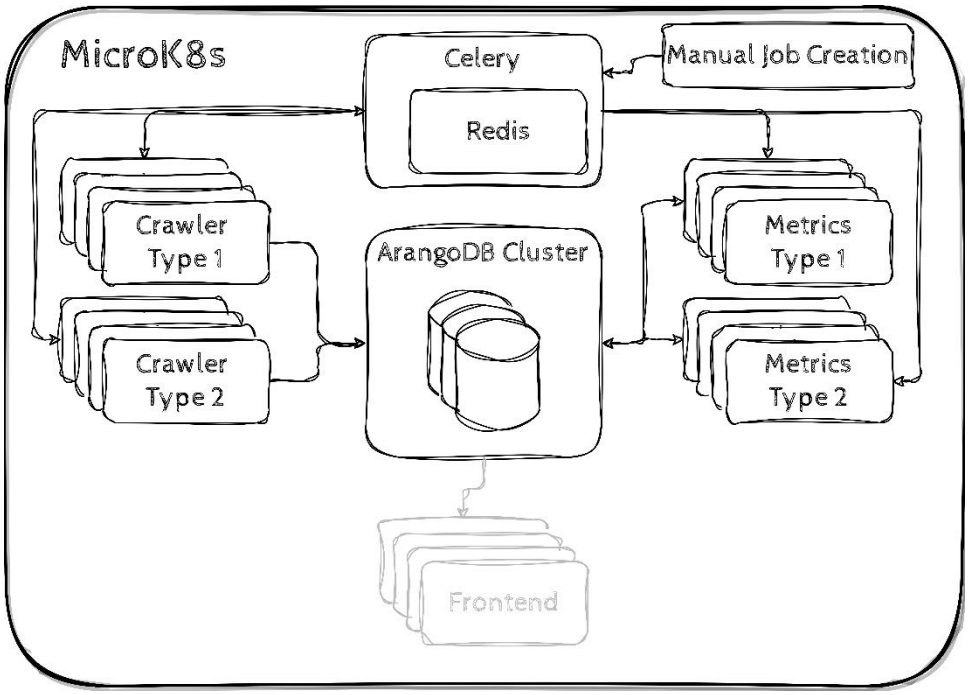
2.4 Arbeitspaket 4 – Implementierung und Crawling

Im Rahmen der ersten Planungen wurde bereits ein abstraktes Architekturdiagramm erstellt:



Gemäß unserem ursprünglich geplanten Aufbau werden Informationen zu Open-Source-Repositorys von verschiedenen Quellen abgerufen und gespeichert. Auf diese Informationen werden unsere Metriken angewandt und die Ergebnisse anschließend wiederum in unserer Datenbank gespeichert. Diese Vorgänge erfolgen regelmäßig, um laufend aktuelle Ergebnisse präsentieren zu können. Diese Ergebnisse sind anschließend über ein Web-Interface beziehungsweise über andere Integrationen abrufbar.

In diesem Arbeitspaket wurde die ursprüngliche Architektur konkretisiert und die einzelnen Komponenten implementiert. Die Architektur sieht nun wie folgt aus:



Unser System wurde aufgrund benötigter Skalierbarkeit und Portabilität als Kubernetes-Cluster konzipiert. Die einzelnen Komponenten, insbesondere die Crawler- und die Metrik-Container, können je nach Bedarf in beliebiger Anzahl eingesetzt werden. Weiters kann das System über mehrere Knoten (Rechner) verteilt werden. Das Sammeln der Informationen zu Repositorys wird mit Hilfe einer Celery Task Queue verwalten. Diese kümmert sich darum, die Scans regelmäßig auszuführen (und im Fehlerfall zu wiederholen) und weist die Aufgabe einem Crawler-Container zu. Es können beliebige Container für das Sammeln von Informationen eingesetzt werden. Aktuell existieren zwei verschiedene Typen: Type 1 stützt sich hauptsächlich auf die GitHub REST API und verwendet die Softwarebibliothek, die von Jacqueline Schmatz im Rahmen ihrer Abschlussarbeit programmiert wurde. Type 2 verwendet großteils die GraphQL-Schnittstelle von GitHub. Die gesammelten Daten werden in einer verteilten ArangoDB-Datenbank gespeichert.

Analog zu den verschiedenen Typen der Crawler können auch verschiedene Typen der Metrik-Berechnung eingesetzt werden. Die Ergebnisse der Metrik-Container werden wiederum in der Datenbank gespeichert. Das Web-Interface lädt die entsprechenden Metriken und stellt diese dann visuell aufbereitet dar. Das Web-Interface wird in AP5 entwickelt werden.

2.5 Arbeitspaket 5 – Plattform, Website und Integration

(Dieses Arbeitspaket wurde noch nicht begonnen.)

2.6 Arbeitspaket 6 – Dokumentation und Formales am Projektende

(Dieses Arbeitspaket wurde noch nicht begonnen.)

3 Umsetzung Förderauflagen

(In der Fördervereinbarung sind keine Förderauflagen vorgesehen.)

4 Zusammenfassung Planaktualisierung

(Wir hatten bislang noch keinen Bedarf für Anpassungen im Projektplan.)

5 Öffentlichkeitsarbeit/ Vernetzung

Die Aktivitäten zur Öffentlichkeitsarbeit und Vernetzung sind teilweise bereits weiter oben in Arbeitspaket 2 beschrieben, hier aber noch einmal eine kurze Aufstellung:

- drei Blogposts im netidee-Blog;
- zwei wissenschaftliche Publikationen (eine dritte in Vorbereitung); damit zusammenhängend
- zwei Präsentationen auf Konferenzen;
- Social-Media-Aktivitäten auf den Accounts der Projektmitarbeiter; sowie
- Berichte im Newsroom der FH St. Pölten und über die Austria Presse Agentur:
    - https://www.fhstp.ac.at/de/newsroom/news/projektfoerderung-fuer-netidee-projekt-crossd
    - https://www.fhstp.ac.at/de/newsroom/news/eine-gesunde-open-source-landschaft-schaffen
    - https://science.apa.at/power-search/15750441504919008005

6 Eigene Projektwebsite

https://crossd.tech

Wie bereits weiter oben in den Aktivitäten zu Arbeitspaket 2 angegeben, wird zu Disseminations-, Outreach- und Networking-Zwecken eine eigene Website unter https://crossd.tech benutzt. Im weiteren Verlauf des Projekts wird diese Webseite auch die Grundlage für die CrOSSD-Plattform sein, d. h. als Benutzeroberfläche für die indizierten Projekte und erfassten Metriken dienen.