Richtlinie zum „Software-Sprint“

Pseudify

Schlussbericht

Zuwendungsempfänger:

Ralf Zimmermann

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS22S44 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Immer mehr Projekte sind heute webbasiert, bereits seit längerer Zeit, aber insbesondere im Rahmen der Verstärkung der Digitalisierung sind mehr und mehr sensible Projekte online. Diese Projekte enthalten personenbezogenen Daten, wie Kundendaten, Adressen, Telefonnummern oder Gesundheitsdaten. Während im Idealfall diese Daten in getrennten Datenbanken vorgehalten werden, ist in der Realität vielfach eine Mischung der Daten in der Applikationsdatenbank vorhanden.

Entwickler verwenden lokale Entwicklungssysteme zum Nachstellen von Fehlern, zur Bearbeitung, Weiterentwicklung und Qualitätssicherung der Produktionssysteme. Damit der lokale Entwicklungsstand möglichst nah an den Produktivstand herankommt, verwenden Entwickler häufig Datenbank-Spiegelungen der Produktionsdaten auf dem lokalen System. Entwicklerrechner stellen damit eine signifikante Gefahr für die Datensicherheit dar. Im Idealfall würden die Entwickler mit anonymisierten oder pseudonymisierten Daten arbeiten, da Kunden-Daten, die nicht vorhanden sind, auch nicht wegkommen können. Aktuell benutzen sicherheitssensitive Entwickler eigene Lösungen – oder keine.

Mit Pseudify wird eine Anwendung geschaffen die:

- Betriebssystemagnostisch funktioniert, sodass Pseudify in möglichst vielen Umgebungen lauffähig ist
- Datenbankagnostisch funktioniert, sodass möglichst viele Applikationen unterstützt werden können
- Applikationsagnostisch funktioniert, sodass Datenbanken von beliebigen Applikationen pseudonymisiert werden können
- es ermöglicht enkodierte Daten in Datenbanken zu pseudonymisieren

- es ermöglicht komplexe Datenstrukturen in Datenbanken zu pseudonymisieren

Pseudify soll PHP basiert sein, da ein Großteil der Websites (je nach Quelle ca. 80 %) PHP verwendet und damit die Einstiegshürde für die Verwendung und Contribution niedrig ist. Erprobte OpenSource-Framework Komponenten (wie zum Beispiel aus dem Symfony Framework oder dem PHP-Faker-Projekt) helfen bei der Entwicklung.

Zudem soll das Projekt in Form von Docker-Containern zur Verfügung gestellt werden, um den Einsatz auf beliebigen Systemen möglichst nahtlos zu ermöglichen – dies ist insbesondere im CI-Kontext relevant.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Die Zielgruppe für das Projekt sind Web-Anwendungsentwickler, vor allem mit der Rolle DevOps, sowie Administratoren von CI-Infrastrukturen.
Mit Pseudify wurde eine Komponente geschaffen, die dazu beitragen kann, das Internet ein Stück sicherer zu machen. Verantwortliche Entwickler und Administratoren von technischen Dienstleistern werden mit Pseudify in die Lage versetzt, Daten aus Datenbanken von Produktionssystemen pseudonymisieren zu können, bevor diese Datenbanken an Entwicklungs- und Testsysteme verteilt werden.
Pseudify bietet dafür neben einem technischen Handwerkskasten auch eine umfängliche mehrsprachige Dokumentation und einen ausführlichen Artikel zum Thema personenbezogene sowie personenbeziehbare Daten.
Zusätzlich liefert Pseudify Vorlagen für Pseudonymisierungsprofile für verschiedene große Webanwendungen wie Wordpress oder TYPO3, die von der Nutzergemeinschaft beständig weiterentwickelt werden können.

Ausführliche Darstellung der Ergebnisse

Mit Pseudify wurde ein Open-Source-Programm geschaffen, dass es Entwicklern ermöglicht, einfach und transparent Datenbank-Spiegelungen von beliebigen Datenbanken mit pseudonymisierten Daten zu erstellen, welches auf Continous-Integration-Servern oder im Rahmen der Datensynchronisation genutzt werden kann.

Pseudify ist Applikationsagnostisch. Das bedeutet, dass beliebige Anwendungsdaten pseudonymisiert werden können, solange die eingesetzte Datenbank von Pseudify unterstützt wird.

Pseudify unterstützt 7 Datenbankplattformen und ist mit anderen Plattformen erweiterbar.

Innerhalb einer Pseudonymisierung werden für die gleichen Eingangsdaten immer die gleichen Pseudonyme gebildet. Dadurch wird die Datenintegrität über alle Datenbanktabellen hinweg gesichert.

Auch wenn Datenbankdaten in enkodierter Form abgespeichert sind, können sie durch Pseudify pseudonymisiert werden.

Pseudify ermöglicht die Manipulation komplexer Datenstrukturen wie JSON oder serialisierter PHP-Datenstrukturen.

Pseudify ermöglicht es, dass die pseudonymisierten Daten den Originaldaten ähneln. E-Mail-Adressen sehen nach der Pseudonymisierung wie E-Mail-Adressen aus, Benutzernamen nach der Pseudonymisierung wie Benutzernamen usw.

Dies stellt sicher, dass die Applikation nach der Pseudonymisierung der Applikationsdatenbank nach wie vor mit den Daten umgehen kann und sich das System analog zum Produktionssystem verhält - ohne Datenschutzrisiken.

Aktuell werden über 100 Datenformate bereitgestellt und es können beliebig viele eigene Datenformate implementiert werden.

Ein Großteil der Webprojekte basiert auf der Programmiersprache PHP (je nach Quelle ca. 80 %).

Pseudify wurde daher in dieser Programmiersprache entwickelt und die Modellierung der Pseudonymisierungsprofile wird darum in dieser Programmiersprache vorgenommen.

Pseudify wird mittels der Virtualisierungstechnologie "Docker" zur Verfügung gestellt. Das Programm ist somit Betriebssystemunabhängig verwendbar.

Durch die Förderung der Förderinitiative „Software-Sprint“ und der tatkräftigen Unterstützung der Open Knowledge Foundation war es möglich diese Software als Open-Source Komponente zu realisieren und der Gesellschaft zur Verfügung zu stellen. Im Rahmen der Fördersumme und des Förderzeitraums konnten alle geplanten Meilensteine umgesetzt werden. Ohne die Förderinitiative hätte es Pseudify in dieser Form nicht geben können.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Verantwortliche Entwickler und Administratoren von technischen Dienstleistern werden mit Pseudify in die Lage versetzt, Daten aus Datenbanken von Produktionssystemen pseudonymisieren zu können, bevor diese Datenbanken an Entwicklungs- und Testsysteme verteilt werden.
Durch den betriebssystemagnostischen, applikationsagnostischen und datenbankagnostischen Charakter der Anwendung ist Pseudify vielseitig einsetzbar.
Die Analysekomponente von Pseudify kann mit beliebigen Suchdaten versorgt werden. Dadurch ist es möglich auch andere Daten als personenbezogene Daten in Datenbanken aufzufinden.
In zukünftigen Versionen von Pseudify soll die Analysekomponente mit KI gestützten Verfahren ausgebaut werden, um das Suchverfahren noch weiter automatisieren zu können.
Die Manipulationskomponente von Pseudify kann generell dazu verwendet werden, um gezielt Datenstrukturen in Datenbanken manipulieren zu können.
Die Pseudonymisierung von Daten in Datenbanken ist somit nur ein Anwendungsfall von vielen. Pseudify könnte z.B. dazu verwendet werden, um die in Datenbanken gespeicherte Passwörter regelmäßig gegen neue zu tauschen und diese neuen Passwörter an die Benutzer zu verteilen.
Pseudify wird in den nächsten Monaten bei 2 großen technischen Dienstleistern implementiert, um die Software in der Praxis zu testen. In diesem Rahmen werden dann mögliche Nachbesserungen an Pseudify vorgenommen.
Durch die Open-Source-Stellung des Projektes ist es der Gemeinschaft möglich, das Programm kontinuierlich an die Bedürfnisse neuer zu pseudonymisierender Applikationen anzupassen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Die Ziele der ersten Version der Software konnten im Rahmen des Förderzeitraums ohne Hindernisse umgesetzt werden.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Auf der Webseite https://www.pseudify.me finden sich Informationen zur Software und eine umfängliche Dokumentation.

Der Quellcode der Software ist auf GitHub unter der Url https://github.com/waldhacker/pseudify-core verfügbar.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Im Projektverlauf ergaben sich keine Planänderungen. Die Software konnte, begründet durch eine genaue Vorplanung, im Rahmen der kalkulierten Stunden umgesetzt werden.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Während des Projektverlaufs konnten keine Entwicklungen anderer Stellen festgestellt werden, die einen Einfluss auf das Projekt oder dessen Zielsetzung zur Folge gehabt hätten.