Generische Pseudonymisierung

als Modul des Zentralen Datenmanagements medizinischer Forschungsdaten

Geidel L, Bahls T, Hoffmann W

Institut für Community Medicine der Universitätsmedizin Greifswald, Abt. Versorgungsepidemiologie und Community Health

Hintergrund

Im Rahmen des GANI_MED-Projektes wurde eine Forschungsplattform konzipiert und umgesetzt, welche unter anderem Funktionen einer Treuhandstelle sowie einer Transferstelle implementiert [1]. Dabei dient die Treuhandstelle der datenschutzkonformen Verarbeitung der personenidentifizierenden Daten (siehe Abbildung 1). Eine ihrer Kernfunktionen ist die Pseudonymisierung aller direkt auf eine natürliche Person rückführbaren Daten wie bspw. eines personenbezogenen Identifiers aber auch Fallnummern oder Laborauftragsnummern. Für die Umsetzung einer solchen Pseudonymisierung sind weitere Anforderungen wie eine mögliche Zweitpseudonymisierung oder der benötigte Wertebereich der Pseudonyme, welcher sich aus Länge und Symbolmenge bestimmt, sowie eine Fehlererkennung durch Prüfzeichen zu berücksichtigen.

Methoden

Es existieren zwei grundsätzliche Methoden, um eine umkehrbare Pseudonymisierung vorzunehmen: die Zuordnung eines Kennzeichens zu einem Pseudonym entweder über einen symmetrischen Algorithmus oder eine arbiträre Zuordnungstabelle [2]. Der algorithmusbasierte Ansatz hat eine Reihe restriktiver Voraussetzungen sowie sicherheitsrelevanter Implikationen [3], daher wurde eine Umsetzung mittels wahlfreier Zuordnungstabelle favorisiert.
Für die zur Fehlererkennung notwendigen Prüfzeichen existieren Standards. Im Besonderen bieten sich mathematische Algorithmen über endlichen Körpern an, wie etwa Derivate des Reed-Solomon-Codes (welche unter anderem für die Fehlerkorrektur bei Audio-CDs eingesetzt werden).

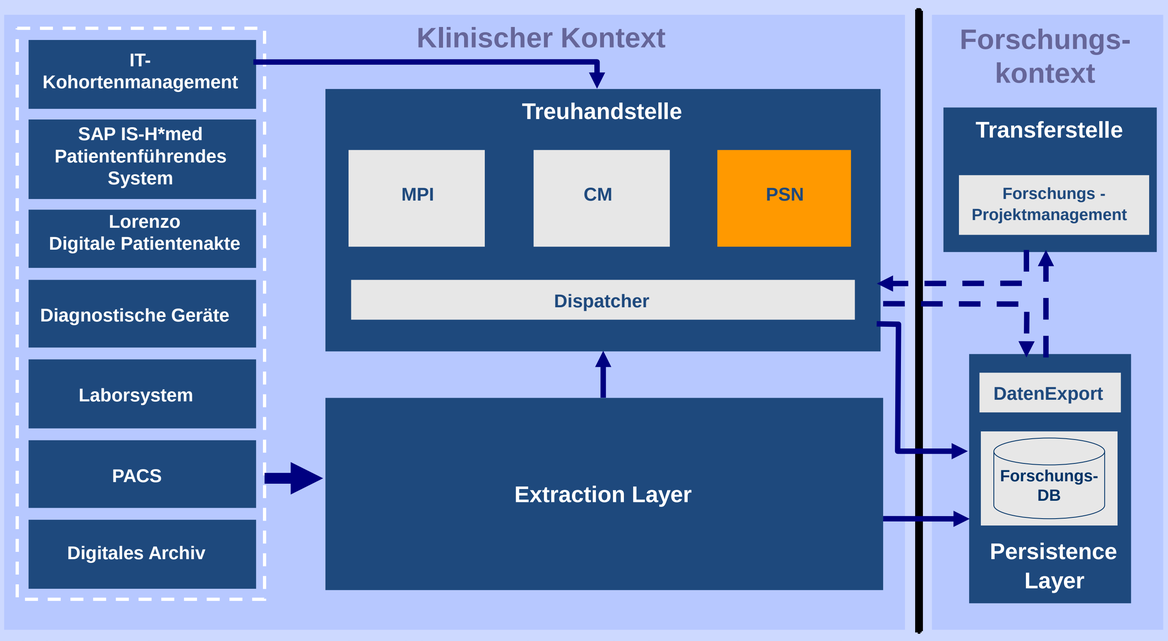


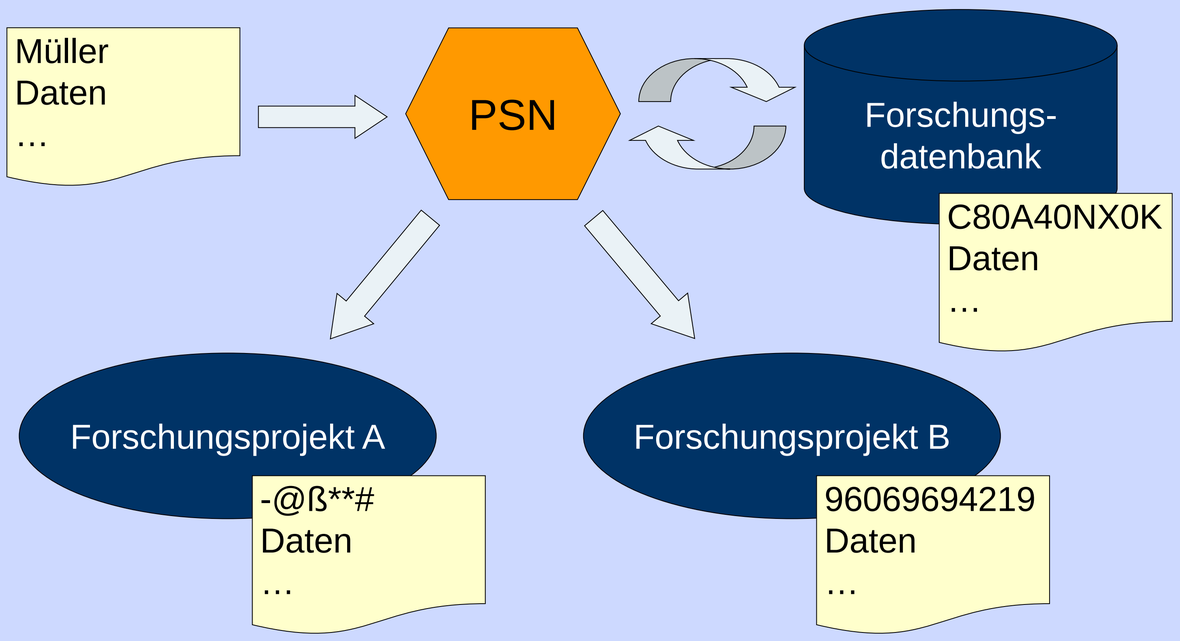
Abbildung 1: Einordnung des Pseudonymisierungswerkzeuges in eine Forschungsplattform

Ergebnisse

Etablierte Werkzeuge, wie der PID-Generator der TMF, vereinen die Funktionalität eines Patientenindizes mit der eines Pseudonymiserungs-moduls [4]. Diese sind fest verbunden und weder einzeln nutzbar noch konfigurierbar. Ferner adressieren diese Implementierungen die Pseudonymisierung eines demographischen Datensatzes, sind jedoch nicht geeignet, um beliebige Daten (zum Beispiel Fall- oder Labornummer) zu verarbeiten. Daher erschienen Konzeption und Implementierung eines generalisierten Moduls als sinnvollere Alternative. Als Designziel wurde eine möglichst umfassende Konfigurierbarkeit etabliert, so dass nun beispielsweise die Symbolmenge für die Pseudonyme in bestimmten Grenzen frei definiert und aus verschiedenen Prüfzeichenalgorithmen ausgewählt werden kann.
Das implementierte Domänenkonzept erlaubt es, mehrere Konfigurationen parallel zu verwenden (dazu zählen die Erzeugung von Pseudonymen 2. Stufe bei Datentransfers oder als pseudonyme Identifier für externe Systeme wie Biobanken).

Abbildung 2 zeigt exemplarisch die Anwendung des generischen Pseudonymisierungswerkzeuges:
Im Rahmen einer Studie erhobene Daten werden mit einem Pseudonym 1. Stufe in einer Forschungsdatenbank gespeichert und, jeweils mit einem Pseudonym 2. Stufe versehen, an verschiedene Forschungsprojekte herausgegeben. Die unterschiedlichen Zweitpseudonyme verhindern dabei eine nicht legitimierte Aggregation herausgegebener Daten.

Domäne	Alphabet	Generatorklasse	Einstellungen
GANI_MED	Symbol31	ReedSolomonLagrange	max_detected_errors=2; psn_length=8
Forschungsprojekt A	a, *, +, -, #, ß, @	ReedSolomonLagrange	max_detected_errors=1; psn_length=5
Forschungsprojekt B	Numbers	Verhoeff	psn_length=10



Originaler Wert	Domäne	Pseudonym
Müller	GANI_MED	C80A40NX0K
C80A40NX0K	Forschungsprojekt A	-@ß**#
C80A40NX0K	Forschungsprojekt B	96069694219

Abbildung 2: Exemplarischer Workflow der Pseudonymisierung; die obere Tabelle zeigt die verwendete Konfiguration, die untere die gespeicherten Pseudonyme

Schlussfolgerungen

Es wurde ein generisches, in weiten Teilen konfigurierbares Modul zur Pseudonymisierung konzipiert und umgesetzt, das aktuellen Anforderungen an den Datenschutz genügt und bereits in mehreren Projekten (GANI_MED, ZDM des DZHK, Zentrales Klinisches Krebsregister Mecklenburg-Vorpommern) erfolgreich eingesetzt wird. Eine effektive Fehlererkennung auf Basis von Prüfzeichen ist unter anderem mittels eines Derivates des Reed-Solomon-Algorithmus gegeben. Die Einbindung in weiterer Projekte ist aufgrund des flexiblen Designs und der Verwendung offener Schnittstellen mit geringem Anpassungsaufwand möglich. Der wissenschaftlichen Gemeinschaft wird das Pseudonymisierungswerkzeug im Rahmen des DFG-Einzelförderprojektes MOSAIC zur Nachnutzung zur Verfügung gestellt (www.mosaic-greifswald.de).

Kontakt

Dipl. Inf Lars Geidel
Institut für Community Medicine, Abt. Versorgungsepidemiologie und Community Health
Universitätsmedizin Greifswald, Ellernholzstrasse 1-2, 17487 Greifswald
Tel.: 03834 / 86-7774, Fax: 03834 / 86-7752, E-Mail: lars.geidel@uni-greifswald.de

Literatur

[1] www.gani-med.de
[2] Pommerening K: Pseudonyme-ein Kompromiß zwischen Anonymisierung und Personenbezug. In: Trampisch HJ, Lange S, Hrsg. Medizinische Forschung - Ärztliches Handeln, 40. München: MMV Medizin-Verlag; 1995. p. 329-33.
[3] Pommerening K, Reng M, Debold P, Semler SC: Pseudonymisierung in der medizinischen Forschung – Das generische TMF-Datenschutzkonzept; in: GMS Med Inform Biom Epidemiol., egms, Köln 2005, Bd 1, Heft 2, Doc17
[4] Reng M, Debold P, Specker C, Pommerening K: Generische Lösungen zum Datenschutz für die Forschungsnetze in der Medizin, Schriftenreihe der TMF – Bd. 1, Berlin, 2006