Der Mainzer PID-Generator

Produktbeschreibung - Stand Juni 2005

Der PID-Generator dient dazu, eine definierte Population mit pseudonymen (nicht-sprechenden) Identifikatoren zu versehen. Dazu wird jeweils Satz von personenidentifizierenden Daten in einen PersonenIDentifikator (PID) transformiert. Berücksichtigt wird dabei, dass die personenidentifizierenden Merkmale fehlerbehaftet und zeitlich veränderlich sein können. Je nach organisatorischen Rahmenbedingungen kann der PID als Pseudonym dienen oder durch eine zusätzliche kryptographische Operation in ein Pseudonym verwandelt werden.

Der Kern des PID-Generators ist eine Datenbank, die die bisher eingegebenen Fälle - die personenidentifizierenden Daten im Klartext oder in umkehrbar oder unumkehrbar verschlüsselter Form - zusammen mit dem jeweiligen PID speichert; wird ein neuer Fall eingegeben, wird wird durch einen Abgleichs- (Match-) Algorithmus festgestellt, ob dieser Fall schon erfasst ist und der bereits gespeicherte PID auszugeben ist oder ob ein neuer PID erzeugt werden muss.

Betriebsmodi

Der PID-Generator ist als Web-Service konzipiert. Dazu wird das Programm über die CGI-Schnittstelle eines Webservers angesprochen; als Benutzungsoberfläche dienen konfigurierbare HTML-Seiten. Darüber hinaus ist ein interaktiver Konsolenbetrieb sowie ein Batch-Betrieb möglich. Im letzteren Fall ist die Ergebnis-Rückmeldung in eine Datei möglich. Für administrative Datenpflege, etwa bei nachträglichem Erkennen eines Homonyms, ist ein Direktzugriff auf die Datenbank zu nutzen. Eine Einbindung in bestehende RDE-Systeme wird angestrebt; eine SOAP-Schnittstelle zu diesem Zweck ist als externes Modul vorhanden.

Konfigurationsdatei

In der Konfigurationsdatei können u. a. definiert werden:

- Struktur der Eingabedatensätze (Feldnamen, Typ, Feldgrenzen, Transformationsoptionen, Abgleichsoptionen),
- Ablauf des Match-Verfahrens,
- Datenbank-Verbindung,
- Log-Datei und Umfang der Log-Aktivitäten,
- Nutzung eines Verzeichnisdienstes über LDAP,
- Templates für die HTML-Seiten, die die Benutzungsoberfläche ausmachen,
- Meldungstexte für die verschiedenen Ausgänge des Match-Verfahrens,
- Schlüssel für die verschiedenen kryptographischen Operationen.

Das Datenbankschema wird aus der Konfigurationsdatei erzeugt.

Match-Verfahren

Dafür liegt eine gesonderte Beschreibung vor. Eine besondere Rolle spielen dabei

Krankenkassen-Nummern, die, falls vorhanden, bei exakter Übereinstimmung stets zu einem Match führen. Durch die Festlegungen in der Konfigurationsdatei ist das Match-Verfahren sehr flexibel konfigurierbar. Eine Änderung des Match-Verfahrens im laufenden Betrieb ist jederzeit möglich.

PID-Erzeugung

Der PID wird als kryptographisch verschlüsselte laufende Nummer erzeugt und mit Prüfzeichen versehen. Auch dafür liegt eine gesonderte Beschreibung vor. Der PID-Algorithmus hängt von einem Schlüssel ab, der in jedem Anwendungsumfeld gesondert gesetzt werden sollte, um die Entstehung eines weit verbreiteten Personenkennzeichens zu verhindern. Im laufenden Betrieb darf der PID-Algorithmus samt seinem Schlüssel *nicht mehr* geändert werden.

Portabilität und Installation

Der PID-Generator ist in reinem C entwickelt und sollte sich auf jedem UNIX-System problemlos kompilieren lassen (getestet unter verschiedenen Linux-Versionen und OpenBSD). Für MS-Windows-Systeme ist eine installationsfertige Version vorhanden. Als Datenbank wird PostgreSQL direkt unterstützt; andere Datenbanken können über eine ODBC-Schnittstelle angebunden werden.

Verwendete Fremdsoftware

Hash-Algorithmus MD5, Verschlüsselungsverfahren AES, Kölner Phonetik, Hannoveraner Phonetik. Einige der Verfahren (Normalisierung von Namen, Erzeugung von Kontrollnummern) sind Weiterentwicklungen von Algorithmen des Landeskrebsregisters Rheinland-Pfalz.

Dokumentation

Ein Handbuch zu Installation, Konfiguration und Betrieb liegt vor.

Geplante Erweiterungen

RDE-Systeme: Der PID-Generator soll online auch aus bestehenden RDE-Systemen aufgerufen werden können. Geplant ist eine exemplarische Realisierung für ein häufig genutztes System.

Stochastisches Match-Verfahren: Das Match-Verfahren wird um eine optimierte stochastische Komponente erweitert, die bei der Entscheidung in Zweifelsfällen auch die Häufigkeiten von Merkmalen berücksichtigt (Fellegi-Sunter-Verfahren).