Zusammenfassung

Workshop ID-Tools
am 10.05.2012
Uhrzeit: 11:00 – 1:00 Uhr
Ort: Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Institut für Med. Biometrie, Epidemiologie und Informatik (IMBEI), Obere Zahlbacher Str. 69, 55131 Mainz

Teilnehmer (39)

Antony, Gisela	Kompetenznetz Multiple Sklerose
Berliner, Sascha	ZeBanC
Borg, Andreas	Universitätsmedizin Mainz, IMBEI
Christoph, Jan	Universität Erlangen-Nürnberg
Döllinger, Christoph	Universitätsklinikum Heidelberg
Drepper, Johannes	Geschäftsstelle TMF e. V.
Eder, Johann	Alpen-Adria Universität Klagenfurt
Enders, Frank	Averbis GmbH
Engel, Marcel W.	ZSE Freiburg
Engelmann, Tobias	IZKS Mainz
Finkenwirth, Thomas	Fraunhofer IBMT
Ganslandt, Thomas	Uniklinikum Erlangen
Geiger, Jörg	Universitätsklinikum Würzburg
Grenz, Michael	Universität Göttingen
Hahmann, Maik	KKS Marburg
Kohlmayer, Florian	Klinikum rechts der Isar der TU-München
Kuchinke, Wolfgang	Heinrich-Heine Univerität Düsseldorf
Lablans, Martin	Universitätsmedizin Mainz, IMBEI
Langner, Dirk	Uni-Greifswald, ICM-VC
Liebmann, Robert	ZSE-Freiburg Klinikum
Löbe, Matthias	Universität Leipzig, Institut für Medizinische Informatik, Statistik und Epidemiologie
Meisner, Christoph	Institut für Medizische Biometrie
Nasseh, Daniel	IBE
Ploetz, Cathleen	Geschäftsstelle TMF e.V.
Pommerening, Klaus	Universitätsmedizin Mainz, IMBEI
Quade, Matthias	Universitätsmedizin Göttingen, Abteilung Medizinische Informatik
Rock, Hans-W.	KNP CIO MR
Schack, Christian	Universitätsmedizin Greifswald ICM-VC
Schepers, Josef	Charité / BFG-Projekt
Schmidt, Robert	RWTH Aachen
Schwanke, Jens	Universitätsmedizin Göttingen
Speer, Ronald	ZKS Leipzig
Stenzhorn, Holger	Universitätsklinikum des Saarlandes
Stropp, Udo	UK Essen Westdeutsches Tumorzentrum
Troska, Siegfried	Biobank popgen
Ückert, Frank	Universitätsmedizin Mainz, IMBEI

Zusammenfassung

Workshop ID-Tools
am 10.05.2012
Uhrzeit: 11:00 – 1:00 Uhr
Ort: Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Institut für Med. Biometrie, Epidemiologie und Informatik (IMBEI), Obere Zahlbacher Str. 69, 55131 Mainz

Volk, Daniel	IZKS Mainz
Werberger, Sven	Institute für Klinische Molekularbiologie
Zerbe, Norman	ZeBanC

Einführung

Ziel dieses Workshops war es, das Leistungsspektrum der bestehenden „ID-Tools“-Werkzeuge kritisch zu reflektieren und daraus zum einen den konkreten Überarbeitungsbedarf der bestehenden Werkzeuge sowie zum anderen die Anforderungen an neu zu entwickelnde ID-Tools herauszustellen. Dazu diskutierten die Teilnehmer Anforderungen aus der Überarbeitung der generischen Datenschutzkonzepte der TMF, konkrete Anforderungen und Erfahrungen aus den Forschungsnetzen sowie Wünsche und Anregungen derzeitiger Anwender.

Bewertung der vorhandenen TMF-Tools

Die Berichte der Teilnehmer loben die Stabilität des PID-Generators, stellen aber auch einen technischen Überarbeitungs- und funktionalen Erweiterungsbedarf fest. Die Installation ist mit einem relativ großen Aufwand verbunden, was unter anderem daran liegt, dass die Konfiguration und Installation nur per Kommandozeile möglich und die Dokumentation unvollständig ist. Die Tatsache, dass das Programm eine eigene (PostgreSQL-) Datenbank benötigt, erschwert außerdem die Einbindung in bestehende Systeme. Der Bedarf an dieser Software zeigt sich aber an zahlreichen Anfragen, auch von Institutionen außerhalb der TMF.

Der Einsatz des TMF-Pseudonymisierungsdienstes (zweiter Stufe, „PSD“) wird als derzeit unpraktikabel angesehen. Grund ist neben einem kaum zu leistenden Installationsaufwand die unflexible Einsatzmöglichkeit, z.B. bei der Verbindung mit mehreren Quell- und/oder Zieldatenbanken („eine Smartcard pro Verbund“).

Anforderungen für neue ID-Tools

Neben technischen Schwächen machen auch neue konzeptionelle Anforderungen eine Neu- oder Weiterentwicklung des PID-Generators nötig. Am meisten nachgefragt ist dabei die Möglichkeit, neben dem PID weitere Typen von Identifikatoren zu verwalten, eine Anforderung, die sich z.B. aus veränderten rechtlichen Rahmenbedingungen und neuen Anwendungsfällen (z.B. Biobanken) ergibt. Als weitere Anforderungen sind zu nennen:

- alternative Verfahren des Record Linkage,
- Unterstützung für fremdsprachliche Datensätze (die Phonetik des alten PID-Generators ist auf deutsche Namen optimiert),

Zusammenfassung

Workshop ID-Tools
am 10.05.2012
Uhrzeit: 11:00 – 1:00 Uhr
Ort: Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Institut für Med. Biometrie, Epidemiologie und Informatik (IMBEI), Obere Zahlbacher Str. 69, 55131 Mainz

- Funktionalitäten für den Einsatz bei behandlungnahen Anwendungen (z.B. Verwaltung von temporären IDs für eine Webanwendung),
- einfache Konfigurier- und Wartbarkeit, um den Betrieb durch externe Dienstleister zu erleichtern (→ Oberfläche + Dokumentation für IDAT-Admin),
- eine aktuelle, aber stabile technische Basis (d.h. Programmierframework) und
- bessere Schnittstellen zur Integration in bestehende Systeme.

Stabilität und Performanz des alten PID-Generators sollen dabei erhalten bleiben, insbesondere auch die Möglichkeit, große Datenmengen effizient als Batchlauf zu verarbeiten. Generell ist eine möglichst umfangreiche Kompatibilität zu bestehenden Konfigurationen und Datenbeständen zu gewährleisten. Aus den schlechten Erfahrungen mit der Komplexität des TMF-Pseudonymisierungsdienstes (zweiter Stufe, „PSD“) werden starke, aber einfach anzusprechende Webschnittstellen vorausgesetzt.

Pseudonymisierung im GANI_MED-Projekt

Herr Schack stellt die in Greifswald entstandene GANI_MED-Pseudonymisierungs-Infrastruktur vor, die als Java EE-Anwendung in einem Java EE Application Server auszuführen ist. Einsatzzweck der Infrastruktur ist die Integration der heterogenen GANI_MED-Datenquellen unter Erzeugung eines systemweit eindeutigen Identifikators für alle Teilsysteme der GANI_MED-Forschungsplattform. Nachrichten werden über SOAP-Schnittstellen ausgetauscht.

Die Infrastruktur führt eine eigene Nutzerliste. Die gewählte WS-Security-Implementierung gewährleistet hohe Sicherheit auf Basis zertifikatbasierter Authentifizierung, macht damit jedoch einen Zugriff durch noch unbekannte Drittsysteme wie etwa Webbrowser vglw. schwierig. Die Software wurde nicht mit Blick auf die TMF-Datenschutzkonzepte erstellt, nach deren alten Fassung käme es jedoch Modell B am nächsten. Eine Nutzung in Modell A ist momentan nicht möglich, da temporäre Identifikatoren nicht vorgesehen sind.

Das Mainzer ID-Framework

Herr Lablans stellt das Mainzer ID-Framework vor, das er gerade u.a. mit Herrn Borg in Mainz entwickelt um den Bedarf an der dortigen Verbund-IT zu decken. Zum Einsatz kommt das ID-Framework dort sowohl im Versorgungsmodul, im Biobanken-Modul als auch im Forschungsmodul.

Ein Java-Servlet bietet eine REST-artige Webschnittstelle an, über die mithilfe einfacher HTTP-Befehle Patienten angelegt werden können. Verwendete Matchfelder und –verfahren sind ebenso parametrier- und erweiterbar wie zu speichernde IDs und ihre jeweiligen Generatoren. Als primäres ID-Verfahren konnte der PID-Erzeugungs-Code portiert werden und sich in automatisierten Tests bewähren. Als primäres Matchverfahren kommt zurzeit ein Verfahren zum Einsatz, welches sich in

Zusammenfassung

Workshop ID-Tools
am 10.05.2012
Uhrzeit: 11:00 – 1:00 Uhr
Ort: Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Institut für Med. Biometrie, Epidemiologie und Informatik (IMBEI), Obere Zahlbacher Str. 69, 55131 Mainz

der Record Linkage-Software Epilink[1] bewährt hat. In Verbindung mit der Verwendung von Bloomfiltern und dem Dice-Koeffizienten als Ähnlichkeitsmaß[2] ermöglicht dieses gewichtsbasierte Verfahren fehlertolerante Vergleiche auf Hashes.

Die Mainzer Patientenliste soll als erster Baustein des Mainzer ID-Frameworks noch im Mai/Juni erstmals produktiv eingesetzt werden. Im ersten Quartal 2013 ist eine allgemeine Freigabe der Software unter einer Open-Source-Lizenz geplant.

Diskussion über Schnittstellen und Funktionalität

Im ursprünglichen Projektantrag war eine SOAP-Schnittstelle für die Patientenliste vorgesehen, wie sie von GANI_MED eingesetzt und sogar zur E-PIX-Referenzimplementierung gebracht wurde. Die Mainzer Patientenliste hingegen wählt stattdessen das REST-Paradigma. Als Grund für diese Entscheidung führt Lablans an, es sei natürlich eine Patientenliste nicht als Nachricht, sondern als Ressource zu begreifen – folglich sei die Entscheidung auf eine ressourcenorientierte Architektur (ROA) auf REST-Basis, nicht auf das nachrichtenorientierte, hier unverhältnismäßig komplexe SOAP gefallen. REST sei außerdem, anders als SOAP, im Webbrowser hervorragend anzusprechen und eigne sich daher auch für Pseudonymisierung in behandlungsnahem Kontext.

Umstritten ist, ob eine Patientenliste selbst authentifizieren sollte (etwa über Zertifikate oder Nutzerkonten). Im Fall der Mainzer Patientenliste wurde stattdessen ein Ticketsystem implementiert.

Anpassung an Verbundbedürfnisse

Die Mainzer Patientenliste, die in Kürze die nötige Kernfunktionalität zur Pseudonymisierung in medizinischen Forschungsnetzen verfügen wird, soll anschließend im Rahmen von TMF-Anträgen weiter auf die spezifischen Bedürfnisse der TMF-Mitglieder angepasst werden. Komplexe Anonymisierungsfunktionen (z.B. zur k-Anonymisierung) sollen dabei zunächst nicht berücksichtigt werden, da hierfür bereits andere Projektaktivitäten in Vorbereitung sind.

Auch nach ihrer Erweiterung sollen die zu entwickelnden Tools nicht zu komplex sein. Erwünscht ist ein hoher Grad von Kooperation, um Doppelarbeit und die Entwicklung konkurrierender inkompatibler Systeme zu vermeiden. Aus Sicht der Softwareentwicklung bietet sich daher ein Open-Source-Projekt an.

---

[1] Contiero P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P, Tessandori R. The EpiLink record linkage software: presentation and results of linkage test on cancer registry files. Methods of Information in Medicine. 2005;44(1):66-71.
[2] Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making. 2009, 9:41. http://www.biomedcentral.com/1472-6947/9/41

Zusammenfassung

Workshop ID-Tools
am 10.05.2012
Uhrzeit: 11:00 – 1:00 Uhr
Ort: Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Institut für Med. Biometrie, Epidemiologie und Informatik (IMBEI), Obere Zahlbacher Str. 69, 55131 Mainz

Herr Quade bereitet dafür in Abstimmung mit den interessierten Kollegen aus Mainz, Greifswald und Erlangen einen ersten Entwurf zur Vorstellung und Diskussion im Rahmen der TMF-Sitzungswoche im September samt anschließender Einreichung beim TMF-Vorstand vor.

k-Anonymisierung, l-Diversität und t-Closeness

Herr Eder verdeutlicht, dass eine einfache De-Identifizierung (Entfernung potentiell identifizierender Daten aus den Datenbeständen) bei Anwendungsfällen mit einem vergleichsweise offenen Zugriff nicht mehr ausreicht, um dauerhafte Vertraulichkeit zu gewährleisten. Selbst bei Anfragen, die nur mit Fallzahlen beantwortet werden und bei denen zusätzlich verhindert wird, dass z.B. Fallzahlen unter einem festgelegten Schwellwert zurückgegeben werden, können spezifische Tracking-Anfragen bei geringem Zusatzwissen über zu suchende Patienten oder Datensätze diese in solchen einfach anonymisierten Datenbeständen aufspüren.

Um solche Attacken durch Trackinganfragen zu verhindern, ist u.a. das Verfahren der k-Anonymisierung für solche Datenbestände vorgeschlagen worden. Dieses Verfahren stellt sicher, dass jede Kombination der auch in anderen Datensammlungen vorhandenen Attribute in mindestens k verschiedenen Datensätzen vorkommt. Wenn jedoch eine kritische Information in allen k Datensätzen in identischer Ausprägung vorkommt, z.B. eine Krankheitsbeschreibung, so kann diese von einem Anfrager doch wieder einem bestimmten der Datensätze zugeordnet werden. Um solchen Attacken (Homogenitätsattacke) zu begegnen, muss der Datenbestand auch dem Kriterium der l-Diversität genügen, die eine ausreichende Variabilität kritischer Attribute innerhalb aller k Datensätze nach einer k-Anonymisierung vorschreibt. Weiteren Attacken bei ungenügenden, bzw. unrepräsentativen Verteilungseigenschaften der kritischen Attribute nach Umsetzung einer l-Diversität kann mit dem Konzept der t-Closeness begegnet werden.

Da mit einer zentralen Zusammenführung von medizinischen Daten zu Bioproben multipler Biobanken potentiell umfangreiche sensitive Informationen für eine große Zahl von Forschern bereit gestellt werden, muss hier ein hoher Anonymisierungsgrad umgesetzt werden, der auch aufwändigeren Trackingattacken standhalten kann. Daher sollte eine solche Datensammlung k-anonym umgesetzt werden und perspektivisch auch den Anforderungen der l-Diversität und t-Closeness genügen. Für eine solche Umsetzung erscheint der OpenAnonymizer von Prof. Eder von der Alpen-Adria-Universität Klagenfurt aus Österreich, der als Open-Source-Tool zur Verfügung gestellt wird, eine vielversprechende Grundlage darzustellen. Eine konfigurierbare k-Anonymisierung wird bereits unterstützt, die Berücksichtigung der Kriterien l-Diversität und t-Closeness ist geplant.

Protokoll: Martin Lablans, Andreas Borg, Johannes Drepper