Anwenderhandbuch

E-PIX®

Enterprise Identifier Cross-Referencing

Version 3.0.0 vom 31.10.2022

Herausgeber:

Unabhängige Treuhandstelle der Universitätsmedizin Greifswald

Autor:

Christopher Hampf, M.Sc.

Ellernholzstr. 1-2
17475 Greifswald

Tel. 03834 / 86-7851, Fax: 03834 / 86-6843
E-Mail: christopher.hampf@uni-greifswald.de

Unabhängige Treuhandstelle
UNIVERSITÄTSMEDIZIN GREIFSWALD

Versionierung

Version	Datum	Bearbeitungsart / Betroffene Abschnitte	Bearbeiter
0.9	19.06.2019	Update und Erweiterung der Dokumentation des Mosaic-Projektes	Christopher Hampf
2.9.2	02.04.2020	Aktualisierung auf E-PIX Version 2.9.2 und Erweiterung um Domänen-Konfiguration	Christopher Hampf
2.10.0	11.02.2021	Aktualisierung auf E-PIX Version 2.10	Christopher Hampf
2.12.0	07.07.2021	Aktualisierung auf E-PIX Version 2.12 Ergänzung der Domänen-Konfiguration um Bloomfilter Aktualisierung aller Bilder	Christopher Hampf
2.13.0	31.03.2022	Aktualisierung auf E-PIX Version 2.13 Ergänzung der Domänen-Konfiguration um Dublettenauflösungsbegründungen Ergänzung um Notifications, Authentifizierung und Optimierungen	Christopher Hampf und Martin Bialke
3.0.0	31.10.2022	Aktualisierung auf E-PIX Version 3.0	Christopher Hampf

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

1 Hintergrund

Um beispielsweise medizinische Daten einer Person eindeutig zuordnen zu können, verwenden Einrichtungen wie Kliniken oder Register typischerweise lokal eindeutige Kennungen (sog. Local Identifier). Diese Kennungen haben jedoch nur innerhalb der jeweiligen Domäne (z.B. Klinik) Gültigkeit. Zudem können identifizierende Daten einer Person, wie Name und Geburtsdatum, aus verschiedenen Quellen aufgrund von Schreibfehlern oder zwischenzeitlichen Änderungen voneinander abweichen, so dass eine Zusammenführung von Daten (Record Linkage) gegebenenfalls nicht erfolgen kann. In diesem Fall spricht man von einem Synonymfehler. Derartige Fehler sind in der Regel nur unter Zuhilfenahme weiterer Daten auflösbar. Werden Daten verschiedener Personen fälschlicherweise einer einzigen Person zugeordnet, entsteht ein Homonymfehler. Diese Fehlerform ist fatal und im Nachgang nur mit sehr hohem Aufwand korrigierbar.

Um Forschungsdaten aus mehreren Projekten und Studien zusammenführen und einer einzigen Person zuordnen zu können, ist sowohl ein Record Linkage als auch eine eineindeutige systemweite Kennung erforderlich, der sowohl die identifizierenden Daten (IDAT) einer Person, als auch die einzelnen lokalen Kennungen des Quellsystems (z.B. Labore, Studienzentralen, etc.) zugeordnet sind. Da dies auch bei unvollständigen oder fehlerhaften Personendaten fehlertolerant und nachvollziehbar erfolgen muss, ist ein nachhaltiges ID-Management erforderlich.

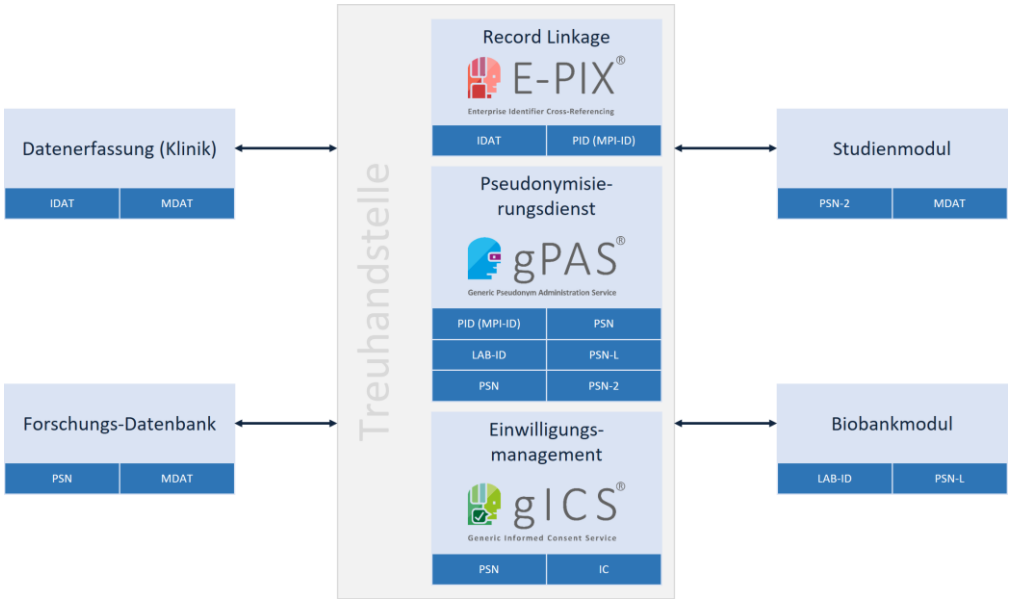


Abbildung 1-1: Das Identitätsdatenmanagement stellt eine zentrale Komponente im medizinischen Forschungskontext dar. Verschiedene Module verwalten modulspezifische Daten und ordnen diese Personen mittels spezifischen Pseudonymen zu. Die Abbildung ist adaptiert vom Maximalmodell des Generischen Datenschutzkonzepts der TMF[1].

[1] POMMERENING, Klaus; HELBING, Krister; GANSLANDT, Thomas; DREPPER, Johannes: Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG, 2014. – ISBN 978–3–95466–123–7

Zweck des ID-Managements ist es, Personendaten unter Vermeidung von Homonymfehlern sicher bereits vorhandenen Datensätzen zuzuordnen und potentielle Dubletten zu erkennen und zusammen zu führen. Ergebnis dieser Zuordnung ist eine systemübergreifende eineindeutige Kennung. Diese stellt gemäß den Konzepten der TMF ein Pseudonym erster Stufe dar. (Quelle: TMF 2004, https://www.tmf-ev.de/Themen/Projekte/V015_01_PID_Generator.aspx, Stand: 07. Dezember 2015)

In der Abteilung Versorgungsepidemiologie und Community Health des Instituts für Community Medicine der Universitätsmedizin Greifswald wurde hierfür der Webservice E-PIX entwickelt. Der E-PIX ist als Open Source Software lizensiert (AGPLv3) und kostenfrei für kommerzielle und nicht-kommerzielle Zwecke einsetzbar.

2 Der Enterprise Identifier Cross-Referencing (E-PIX)

Der E-PIX-Service (kurz für: Enterprise Identifier Cross-Referencing) setzt das Konzept eines Master Patient Index (MPI) um und stellt die notwendige technische Funktionalität zur eindeutigen Identifizierung von Personen in Form eines Webservices bereit. Frei konfigurierbare Personenattribute, typischerweise Vorname, Nachname, Geburtsdatum, Geschlecht, sind Grundlage für die probabilistischen Verfahren zur Zusammenführung von Datensätzen.

Zur Dublettenerkennung wird ein Algorithmus nach Fellegi-Sunter verwendet. Für den Vergleich von Attributen stehen mehrere Vergleichsfunktionen zur Verfügung. Standardmäßig kommt die Levenshtein-Distanz zum Einsatz. Auf diese Weise kann die Zuordnung von Person und eindeutiger systemübergreifender Kennung auch bei unvollständigen bzw. fehlerhaften demografischen Informationen korrekt erfolgen.

Der E-PIX unterstützt neben den erwähnten Vergleichsfunktionen auf Basis von Personendaten im Klartext auch ein Privacy-Preserving Record Linkage (PPRL). Hierbei werden Personendaten derart codiert, sodass keine Rückschlüsse mehr auf die eigentliche Person gezogen werden kann, jedoch dennoch auf Basis dieser codierten Daten vergleiche durchgeführt werden können.

Der E-PIX ermöglicht außerdem die Speicherung domänenspezifischer Lokaler Identifier und standardisierter IHE-Profile (PIX, PDQ). Zudem setzt der E-PIX das Konzept multipler Personenidentitäten um, d.h. einer real existierenden Person können mehrere Ausprägungen (ähnlicher) demografischer Daten zugeordnet sein. Darüber hinaus wird die Auflösung von Synonymfehlern (s. **Abschnitt 4**) unterstützt.

3 Begriffsbestimmungen

Person

Eine natürliche Person, beschrieben durch eine oder mehrere Personenidentitäten.

Personendaten / Identifizierende Daten (IDAT)

Alle Attribute wie Vorname, Nachname, Kontaktdaten, etc. die einer Person zugeordnet sind. Attribute, die eine Person eindeutig identifizieren, werden als identifizierende Daten bezeichnet.

Personenidentität

Bezeichnet eine konkrete Ausprägung eines IDAT-Satzes einer Person. Eine Person kann mehre Identitäten (Haupt- und Nebenidentitäten) besitzen, die sich zum Beispiel in ihrer Schreibweise oder Aktualität unterscheiden (s. **Abschnitt 4**).

Referenzidentität

Die Referenzidentität ist die Hauptidentität einer Person. Wird beispielsweise nach einer Person gesucht, zeigt der E-PIX diese Referenzidentität in der Ergebnisliste an. Andere Identitäten (Nebenidentitäten/Ausprägungen der IDAT) der Person können in der Detailansicht eingesehen werden.

Identifier einer Personenidentität

Eineindeutiger Identifikator (z.B. eine ID) um eine Personenidentität eindeutig zu identifizieren.

Lokaler Identifier

Ein Lokaler Identifier ist ein Identifikator, der durch ein externes System vergeben wurde, wie beispielsweise einem KIS-System. Der Lokale Identifier identifiziert dabei die Personenidentität eindeutig in diesem System. Aus einem System können dabei mehrere Lokale Identifier stammen (z.B. Patienten-ID und Fallnummer). Der Personenidentifikator (PID) kann in seiner Funktion als Identifier auch als LID ("*Lokaler (externer) Identifier*") betrachtet werden.

Domain (Domäne)

Eine Domain ist eine organisatorische Einheit (Mandant), z.B. eine Studie, ein Projekt oder ein Institut.

Lokale Domain (oder auch Identifier Domain)

Domäne des Lokalen Identifiers. Diese muss nicht dem Quellsystem entsprechen. Aus einem Quellsystem können mehrere Lokale Identifier stammen, bspw. Patienten-ID und Fallnummer aus einem KIS. Gleichzeitig kann die gleiche Lokale ID aus unterschiedlichen Quellen stammen, bspw. eine Fallnummer aus einem elektronischen KIS-Export sowie die gleiche Fallnummer von einem Arztbrief.

Matching-Parameter

Frei wählbares Personenattribut (z.B. Vorname, Nachname, Geburtsdatum, etc.), das für das Matching-Verfahren verwendet werden.

Record Linkage

Verfahren um Datensätze einer Person einander zuzuordnen. Hierzu wird die Ähnlichkeit definierter Personendaten (vgl. Matching-Parameter) ermittelt und bei hinreichender Übereinstimmung ein und derselben Person (als Personenidentität) zugeordnet.

Match-Typen

Das Ergebnis des Record Linkages wird klassifiziert in 4 Typen:

1. Perfect Match: Zwei Datensätze sind identisch und gehören zur selben Person.
2. Automatic/Good Match: Zwei Datensätze sind geringfügig unterschiedlich (z.B. durch einen Tippfehler) und werden automatisch derselben Person zugeordnet.
3. Possible Match: Zwei Datensätze weisen Unterschiede auf, sind aber dennoch ähnlich genug, umderselben Person zugehörig sein zu können. Dies erfordert ggf. eine Korrektur der Daten und ein manuelles Auflösen.
4. No Match: Zwei Datensätze weisen keine oder kaum Ähnlichkeiten auf und gehören unterschiedlichen Personen.

Quelle

Datenquelle, aus der IDAT stammen können, z.B. ein Krankenhaus oder ein Forschungsprojekt. Bei der Registrierung einer Personenidentität wird die Quelle ausgewählt, aus der die jeweiligen IDAT stammen. Es kann eine *Sichere Quelle* einer Domäne zugeordnet werden. IDAT die über die Sichere Quelle registriert werden, werden als korrekte Ausprägung einer Personenidentität angesehen (Hauptidentität, vgl. Personenidentität).

4 Das Konzept der Nebenidentitäten

Vor allem bei epidemiologischen Kohortenstudien ist es oftmals erforderlich, die Variationen von IDAT beispielsweise in Bezug auf die (möglicherweise fehlerhafte) Schreibweise eines Namens: Müller, Mueller, Muller, Mülller, etc. im jeweiligen Quellsystem zu erhalten und dennoch die Datensätze eineindeutig einer real existierenden Person fehlerfrei zuordnen zu können.

Innerhalb des E-PIX kann eine Person mehrere Identitäten besitzen, wovon nur eine als Hauptidentität (vgl. Referenzidentität) deklariert werden kann. Die Hauptidentität wird als "die korrekte Ausprägung" der IDAT angesehen. Jede weitere Ausprägung wird als Nebenidentität gespeichert. Ein nachträgliches Ändern der Identitätenbeziehungen ist problemlos möglich, sollte jedoch nur durch autorisiertes Personal und nach eingehender Recherche der Sachlage erfolgen.

Das Konzept von Haupt- und Nebenidentitäten ist in epidemiologischen Kohortenstudien von besonderer Relevanz und ist gleichzeitig Grundlage für das Beheben möglicher Synonymfehler.

5 Funktionalitäten

5.1 Was leistet der Dienst

- Erstellung und Verwaltung einer systemweit eindeutigen Kennung mittels Indexgenerator nach dem Konzept des Master Person Index
- Zusammenführung von Personendaten aus unterschiedlichen Quellsystemen anhand demographischer Informationen
- Umgang mit fehlerhaften/unvollständigen Personendaten
- Unterstützung bei der Rekontaktierung durch die integrierte Personenverwaltung
- Unterstützung beim Auflösen von möglichen Matches durch das Konzept von Haupt- und Nebenidentitäten
- Unterstützung der IHE-Profile PIX & PDQ (PIX ist derzeit noch ohne Update Notification)
- Protokollierung von Systemprozessen und (kritischen) Systementscheidungen
- Beschleunigtes Matching durch Caching: die für den Matching-Prozess erforderliche Datenbasis wird vollständig im Zwischenspeicher gehalten und erlaubt beispielsweise Antwortzeiten beim Anlegen oder Aktualisieren einer Person und einem Datenbestand von bereits 1.000.000 Personen in deutlich weniger als 1 Sekunde
- Einfache Bedienung durch eine intuitive grafische Oberfläche
- Versenden von Notifications bei Zustandsänderungen, um andere Systeme zu informieren

5.2 Was leistet der Dienst nicht

- Eine automatisierte Transkription und Transliteration von demografischen Informationen sind nicht möglich.
- Die Vergabe von Pseudonymen zweiter und weiterer Stufen findet nicht im E-PIX statt, sondern kann in Kombination mit dem gPAS erzielt werden.

6 Installation per Docker

6.1 Systemanforderungen

Technisch / Infrastruktur

- Installierte aktuelle Version von Docker[2] und Docker-Compose[3]
- Administrative Rechte
- Keine Nutzungsbeschränkungen auf die bereitgestellten Service- und Client-URLs

[2] Weitere Informationen unter https://docs.docker.com/install/
[3] Weitere Informationen unter https://docs.docker.com/compose/install/

- Windows oder Ubuntu Server (oder vergleichbar) mit min. 8 GB Arbeitsspeicher, 5 GB Festplattenspeicher, Prozessor (benötigter Arbeitsspeicher und Prozessor-Leistung sind abhängig von erwarteter Datenmenge und -durchsatz)

Software: Anwendungs- und Datenbankserver (ohne Verwendung von Docker)[4]

- JDK 17 oder höher
- WildFly 26 oder höher
- EclipseLink 2.7.11
- MySQL-Connector 8 oder höher
- MySQL-Server 8 oder höher

Personell

- Mitarbeiter mit grundlegenden IT-Kenntnissen zur Administration des Servers und zur Einrichtung des E-PIX-Dienstes (zuzüglich der Wartung und regelmäßiger Sicherungen der E-PIX-Datenbank)
- Ein autorisierter Verantwortlicher zur Administration der E-PIX-Inhalte inkl. zur Auflösung möglicher Matches nach ausführlicher Prüfung der individuellen Sachlage

6.2 Download und Starten des Dienstes

Um den E-PIX als Docker-Container zu starten, werden die Programme *Docker* und *Docker-Compose* benötigt. Beide Programme müssen hierfür installiert sein. Da zwischen beiden Programmen Inkompatibilitäten auftreten können, wird empfohlen die jeweils aktuellsten Versionen zu installieren.

Der E-PIX benötigt zur Ausführung zwei Container (vgl. Abbildung 6-1). Damit diese nicht einzeln gestartet und entsprechend zusammengeschaltet werden müssen, wird der Dienst mit Docker-Compose gestartet. Die entsprechenden Ressourcen können von der THS-Webseite[5] heruntergeladen werden.

[4] Beim Betrieb unter Windows ist zu beachten, dass bei der Verwendung von Volumes und parallel betriebenen VPN-Clients Probleme auftreten können.
[5] https://www.ths-greifswald.de/forscher/e-pix/ bzw. https://www.ths-greifswald.de/forscher/e-pix/#download

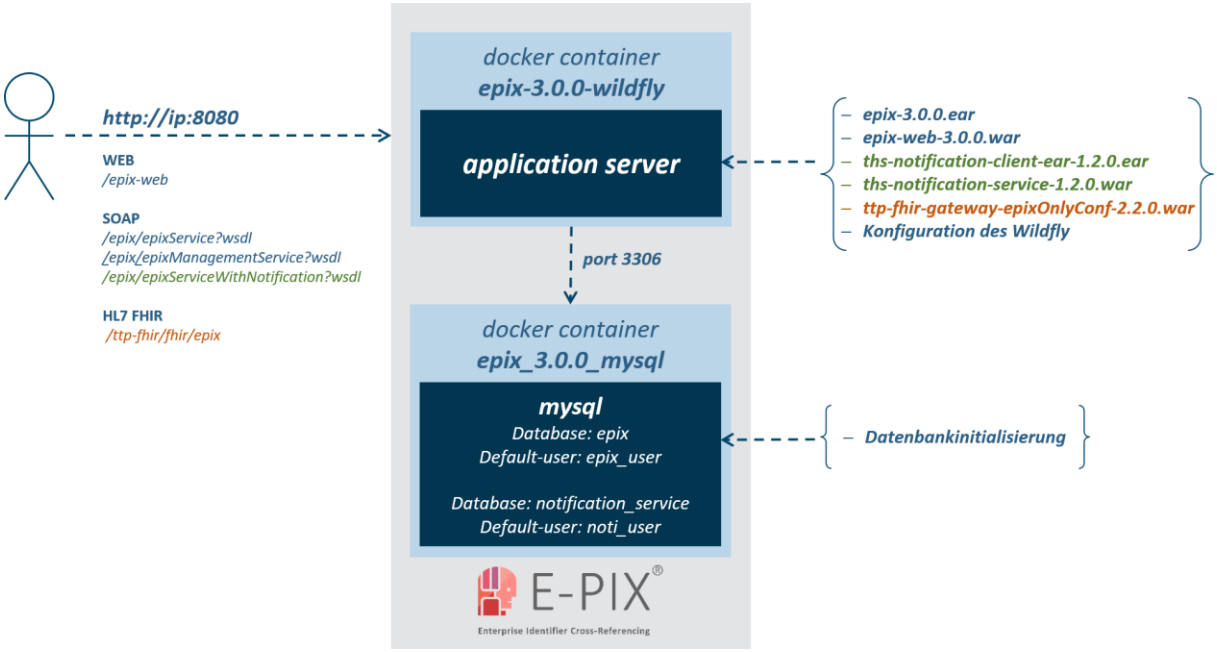


Abbildung 6-1: Architektur des E-PIX® mit Docker.

Das Docker-System besteht aus zwei getrennten Containern. Zum einen aus einer Datenbankinstanz (MySQL) und zum anderen aus dem Anwendungsserver (WildFly inkl. Datenbank-Konnektoren). Der Anwendungsserver kommuniziert mit dem MySQL-Server über den Port 3306. Der Zugriff auf das System von „außen“ erfolgt über den Web-Browser. Die Inhalte werden über den Port 8080 (E-PIX) für den Anwender bereitgestellt.

Um die folgenden Schritte problemlos durchführen zu können, wird ein Account mit administrativen Rechten benötigt. Exemplarisch werden die folgenden Befehle mit **`sudo`** ausgeführt.

Download der benötigten Dateien

Laden Sie die aktuellste Version von https://www.ths-greifswald.de/forscher/e-pix/#download herunter und entpacken Sie die ZIP-Datei. Diese enthält alle relevanten Docker-Compose-Dateien. Im Folgenden wird davon ausgegangen, dass der Ordner in das Verzeichnis `/opt/` entpackt wurde. Der Pfad kann bei Bedarf angepasst werden.

Vergabe von Schreibrechten

```
sudo chmod -R 755 /opt/compose-wildfly/

chown -R 1000:1000 /opt/compose-wildfly/logs/ /opt/compose-
wildfly/deployments/
```

Aus Gründen von Leistung und Ausfallsicherheit sollten die Container des E-PIX auf einem dedizierten Server eingerichtet werden. Zur Administration werden der User `epix` (uid 1000) aus der Gruppe `users` (gid 1000) genutzt.

Wechseln in das E-PIX-Verzeichnis für die Standard-Version

```
cd /opt/compose-wildfly/
```

Starten des E-PIX mithilfe von Docker Compose

```
sudo docker-compose up
```

Damit werden die benötigten Komponenten heruntergeladen[6] und die Konfiguration von MySQL und WildFly gestartet. Danach wird die aktuelle Version des E-PIX bereitgestellt. Der Installationsvorgang kann in Abhängigkeit der vorhandenen Internetverbindung etwa 5 Minuten dauern. Der erfolgreiche Start des Dienstes wird mit der folgenden Ausgabe abgeschlossen.

```
Wildfly 26.1.2.Final […] started in …
```

⚠ **Hinweis:** Weitere Details zur Nutzung von *Docker-Compose* und E-PIX® sind der beigelegten Beschreibung docker-compose/README.md zu entnehmen.

⚠ **Hinweis:** Beachten Sie, dass beim Wechsel vom E-PIX Version 2.x auf 3.0 auch die Docker-Compose Komponenten komplett aktualisiert werden (Umstieg auf Java 17 und WildFly 26). Weitere Hinweise zur Aktualisierung der Docker-Komponenten können der beigelegten Docker-Update.md entnommen werden.

7 Die grafische Benutzeroberfläche des E-PIX

Um dem Treuhandstellenpersonal die Administration der Identitätsdaten zu erleichtern, verfügt der E-PIX über eine grafische Benutzeroberfläche, die speziell für den Einsatz im Web-Browser entwickelt wurde. Der Aufbau der Oberfläche orientiert sich an typischen Arbeitsabläufen innerhalb einer Treuhandstelle.

7.1 Anlegen von Domänen, Quellen und Identifier-Domänen

Der E-PIX erlaubt die Verarbeitung von Personendatensätzen mehrerer Mandanten innerhalb einer Datenbank durch die Verwendung von Domänen (vgl. **Abschnitt 3** Begriffsbestimmungen). Die registrierten Personen sind nur innerhalb einer Domäne eindeutig. Ein Record Linkage findet demnach ebenfalls nur innerhalb einer Domäne statt. Um Personen registrieren zu können, muss eine entsprechende Domäne angelegt werden. Für jede Domäne müssen eine *Sichere Quelle* (vgl. Quelle) und eine Identifier-Domäne angegeben werden. Diese müssen vor dem Anlegen der Domäne im System angelegt werden. Die nötigen Schritte sind unter dem Menüpunkt *Domänen* vorzunehmen und

[6] Sollte Ihre Maschine keinen Zugang zum Internet haben, können die benötigten Images (MySQL und Wildfly) von einer anderen Maschine heruntergeladen werden und dann auf Ihr Zielsystem kopiert werden (siehe https://docs.docker.com/engine/reference/commandline/image_save/ und https://docs.docker.com/engine/reference/commandline/load/).

werden im Folgenden beschrieben. **Abbildung 7-1** zeigt die grafische Oberfläche zum Anlegen von Domänen, Datenquellen und Identifier-Domänen.

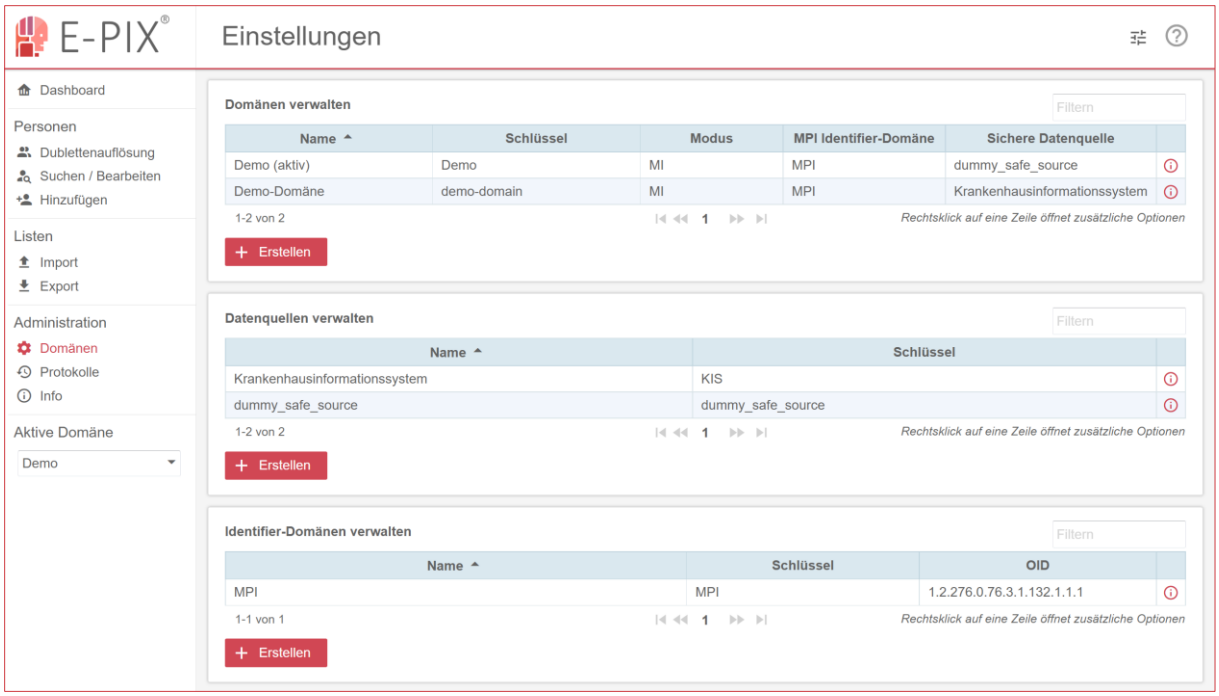


Abbildung 7-1: Oberfläche zum Anlegen von Domänen, Quellen und Identifier-Domänen.

7.1.1 Anlegen einer neuen Quelle

Eine Quelle gibt an, woher die später registrierten Personendaten stammen, also bspw. aus einer bestimmten Studie oder einem Krankenhausinformationssystem (vgl. **Abschnitt 3** Begriffsbestimmungen). Die Quelle kann bei einer Personenregistrierung aus der Liste der zuvor angelegten Einträge ausgewählt werden. Mithilfe der Schaltfläche *Neue Quelle* wird ein neuer Eintrag angelegt. Hierbei kann ein eindeutiger Name und idealer Weise eine Beschreibung angegeben werden. Die *Sichere Quelle* einer Domäne definiert, woher die Hauptidentitäten (respektive die Personendaten, welche als korrekte Ausprägung angesehen werden) stammen.

7.1.2 Anlegen einer Identifier-Domäne

Die Domäne eines Lokalen Identifiers, die so genannte Identifier-Domäne wird auf ähnliche Weise angelegt, wie die Quelle. Hierbei müssen sowohl der Name, als auch die OID eindeutig sein. Jede Forschungseinrichtung besitzt typischerweise eine OID, welche hier angegeben werden kann. Für andere Quellen wie ein KIS, eine Studie etc., kann die OID frei gewählt werden. Wird keine OID angegeben, erzeugt der E-PIX automatisch eine eindeutige Kennung.

7.1.3 Anlegen einer neuen Domäne

Nachdem die *Sichere Quelle* sowie die *Identifier-Domäne* angelegt wurden, kann ein neuer Domänen-Eintrag über die Schaltfläche *Neue Domäne* erstellt werden. Hierfür muss ein eindeutiger Name, die *MPI Domäne* und die *Sichere Quelle* eingetragen werden. Eine Beschreibung sollte insbesondere bei der Verarbeitung von Personen für mehrere Mandanten oder Projekte innerhalb eines E-PIX

eingetragen werden. Die Konfiguration ist im XML-Format vorzunehmen (die Domänenkonfiguration wird im **Kapitel 15** erklärt). Dem E-PIX liegen eine Vielzahl von Konfigurationen bei, welche als Grundlage für projektspezifische Anforderungen angepasst werden können.

7.2 Registrieren einer Person

Bevor eine Person angelegt bzw. registriert werden kann, muss die *Aktive Domäne* ausgewählt werden, für die die Person hinzugefügt wird. Hierzu wird im linken Menü die entsprechende Domäne über das Auswahlmenü gewählt. Wenn nur eine Domäne angelegt wurde, ist diese standardmäßig aktiv. Über den Menüpunkt *Hinzufügen*, wird ein Formular aufgerufen, in welches die Stammdaten/Personendaten eingetragen werden können. Pflichtfelder sind mit einem Stern (*) gekennzeichnet. Welche Felder Pflichtfelder sind, wird in der Konfiguration der Domäne festgelegt (siehe Domänenkonfiguration Pflichtfelder in **Abschnitt 15.4.7**). Es kann zu jeder Person außerdem noch Adress- bzw. Kontaktdaten und beliebig viele Lokale Identifier hinterlegt werden. Weitere Adress- bzw. Kontaktdaten können auf der Detailseite der Person hinzugefügt werden (siehe **Abschnitt 7.4**). Beim Anlegen können Ein- und Auszugsdatum angegeben werden. Die Aktualität einer Adresse kann zusätzlich bearbeitet werden. Mithilfe der Domänen-Konfiguration können noch weitere Felder definiert und benannt werden (vgl. **Abschnitt 15.4.8**). Die *Datenquelle* aus der die Daten stammen muss ebenfalls angegeben werden. Entspricht die angegebene *Datenquelle* der *Sicheren Quelle* der jeweiligen Domäne, dann wird bei Feststellung eines Duplikates die Identität als Hauptidentität deklariert. Diese gilt dann als fehlerfrei (Änderungen und Fehlerkorrekturen können später trotzdem vorgenommen werden. Grundsätzlich kann die Referenzidentität frei gewählt werden). Andernfalls wird eine neue Nebenidentität angelegt. Vor der Registrierung führt der E-PIX ein Record Linkage durch, welcher ermittelt, ob die Person bereits in dieser oder ähnlichen Form hinterlegt ist. Über das Ergebnis dieses Vorgangs informiert der E-PIX. In Abbildung 7-2 wird exemplarisch das Eintragen der Pflichtfelder dargestellt.

ⓘ Was passiert, wenn ein lokaler Identifier bei zwei Identitäten identisch ist?

Wenn die beiden Identitäten zu einem hohen Grad (konfigurationsabhängig) übereinstimmen, dann werden beide Identitäten einer Person zugeordnet. Können die Identitäten nicht einer Person zugeordnet werden, weil keine oder nur eine geringe Übereinstimmung vorliegt, so wird ein Fehler geliefert. Der Grund hierfür ist, dass jeder Identifier nur einer Person zugeordnet sein darf (mehrere Identitäten (Ausprägungen einer Person) können denselben Identifier aufweisen, diese müssen dann aber alle derselben Person zugeordnet sein).

ⓘ Was passiert, wenn zwei Identitäten identisch (*perfect Match*) sind, aber die lokalen Identifier verschieden sind?

Die lokalen Identifier werden der bereits vorhandenen Identität angefügt. Es können mehrere Identifier einer Identität angefügt werden, auch wenn diese aus derselben Identifier-Domäne stammen (Beispiel: Fallnummern). Voraussetzung ist, dass derselbe lokale Identifier niemals unterschiedlichen Personen zugeordnet ist.

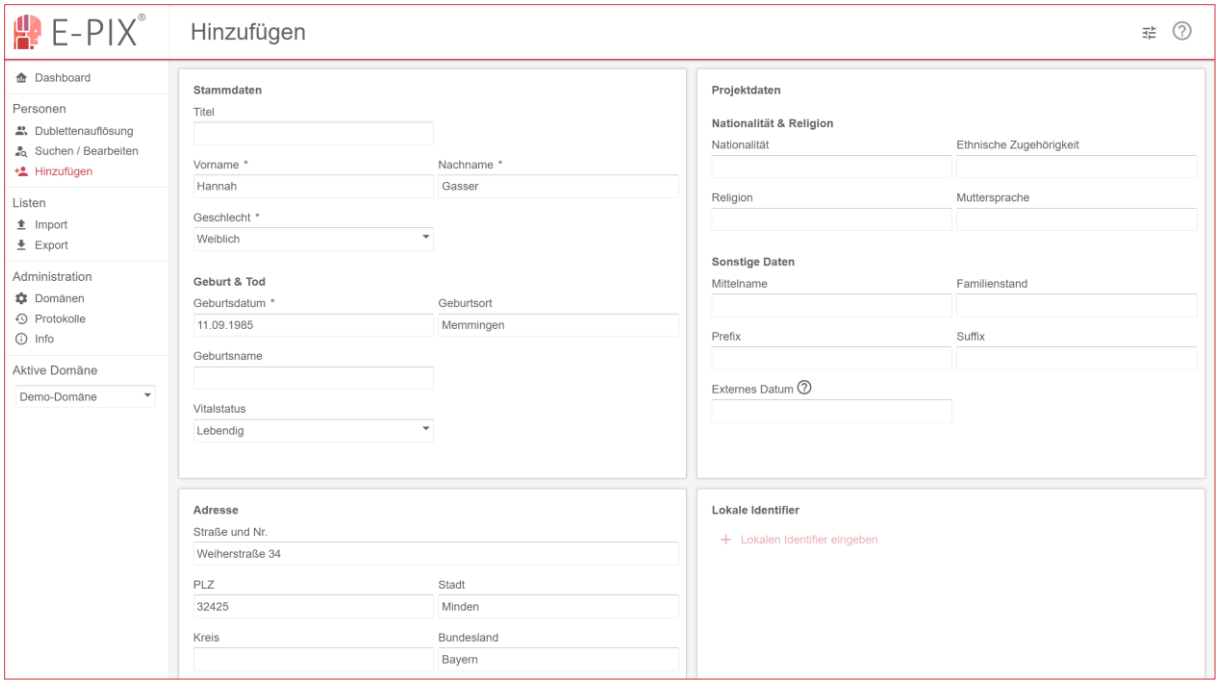


Abbildung 7-2: Oberfläche zum Eintragen von Personendaten.

Record Linkage und Match-Typen

Bei der Registrierung der Person findet ein Abgleich der IDAT statt. Sind diese hinreichend ähnlich zu einer bereits zuvor registrierten Person, so werden diese Personen zusammengeführt. Eine Mitteilung informiert über Erfolg oder Misserfolg. Abhängig von der jeweiligen Domänen-Konfiguration unterscheidet man nach einem Record Linkage unterschiedliche Matchtypen. Diese sind in**Tabelle 7-1** dargestellt.

Tabelle 7-1: Mögliche Match-Typen

Match-Typ	Beschreibung
Perfect Match / Perfekter Match	Exakte Übereinstimmung zweier Datensätze in Bezug auf die Matching-Parameter. Es wird keine neue Person und keine neue Identität angelegt, da die Personendaten bereits in gleicher Form hinterlegt sind.
Good/Automatic Match bzw. Match	Im Hinblick auf den konfigurierten Schwellwert haben zwei Datensätze eine hinreichende Ähnlichkeit. Die neu angegebenen Personendaten werden der bereits bestehenden Person als neue Identität zugeordnet.
Possible Match / Möglicher Match	Es besteht eine Ähnlichkeit zwischen zwei Datensätzen. Bei einem *möglichen Match* findet jedoch keine automatische Zusammenführung statt. Eine Dublettenauflösung kann nur manuell im Nachgang unter Zuhilfenahme weiterer Informationen erfolgen (siehe Anwendungsfall 5 in **Abschnitt 7.6**).

Non-Match / Kein Match	Keine Ähnlichkeit zu einem bestehenden Datensatz. Wenn kein Duplikat festgestellt wurde respektive die Person noch nicht bekannt ist, dann wird eine neue Person angelegt.

7.3 Suchen einer Person anhand demografischer Informationen

Unter dem Menüpunkt *Suchen / Bearbeiten* kann nach Personen gesucht werden, welche mit den angegebenen demographischen Daten übereinstimmen. Neben den Stammdaten/Personendaten kann nach dem MPI oder Lokalen Identifiern gesucht werden. Es müssen hierbei nicht alle Attribute ausgefüllt werden. Die Attribute sind dabei standardmäßig UND-Verknüpft, sodass die Ergebnisliste nur Personen enthält, die alle angegebenen Attribute aufweisen. Alternativ kann auch eine ODER-Verknüpfung erfolgen, sodass die Ergebnisliste nur Personen aufweist, die zumindest mit einem der angegebenen Attribute übereinstimmt. Zum Umschalten ist ein Schalter mit der Bezeichnung *Verknüpfung der Suchparameter* vorhanden. In **Abbildung 7-3** wird exemplarisch eine Person anhand der Attribute Vorname, Nachname und Geschlecht gesucht. Die Ergebnisliste enthält genau einen Eintrag.

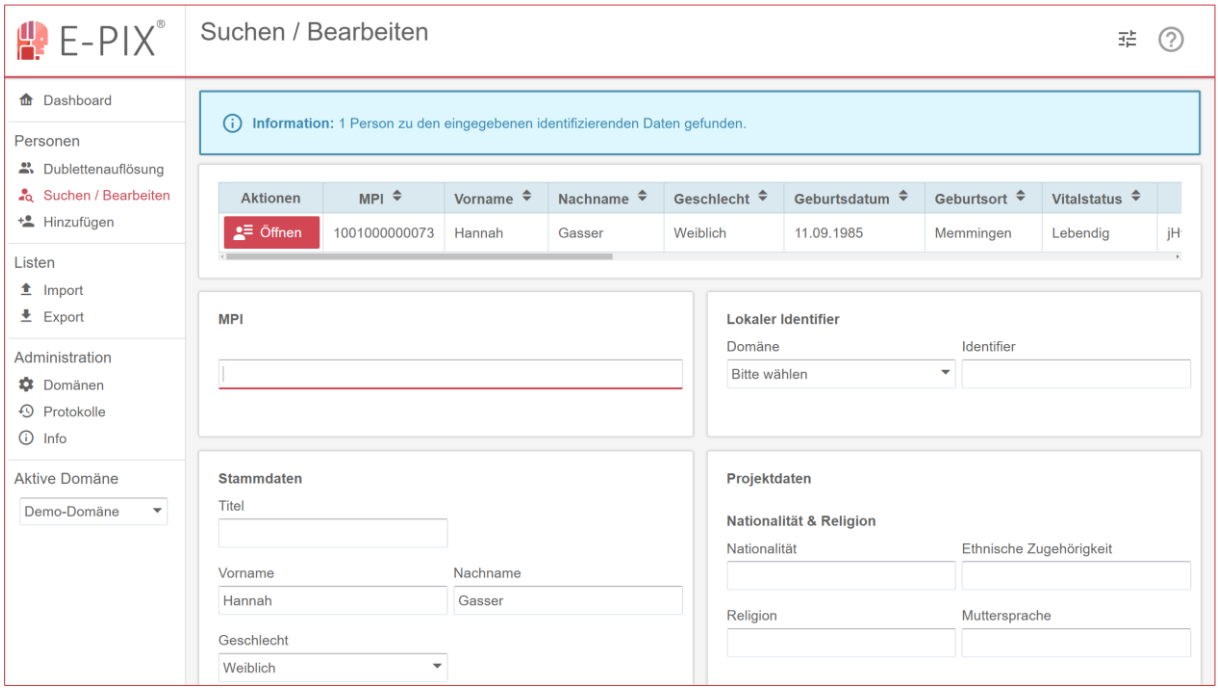


Abbildung 7-3: Oberfläche zum Suchen von Personen anhand von demographischen Daten.

7.4 Einsehen von Details zu einer Person

Um die Detailseite einer Person aufzurufen, muss zunächst nach der betreffenden Person gesucht werden (siehe **Abschnitt 7.3**). In der Ergebnisliste kann über die *Öffnen*-Schaltfläche die Detailseite zur jeweiligen Person aufgerufen werden. Neben den Stammdaten können über die Seite die bekannten Ausprägungen/Identitäten eingesehen werden. Darüber hinaus ist eine Auflistung aller bekannten Adressen vorhanden, sowie ein Zeitstrahl mit allen Änderungen, die diese Person betreffen. Wenn

parallel auch ein gPAS zur Pseudonymverwaltung betrieben wird, kann direkt der Eintrag mit der entsprechenden MPI im gPAS aufgerufen werden. Änderungen werden ebenso über diese Seite durchgeführt. In **Abbildung 7-4** ist exemplarisch die Detailseite einer Person dargestellt.

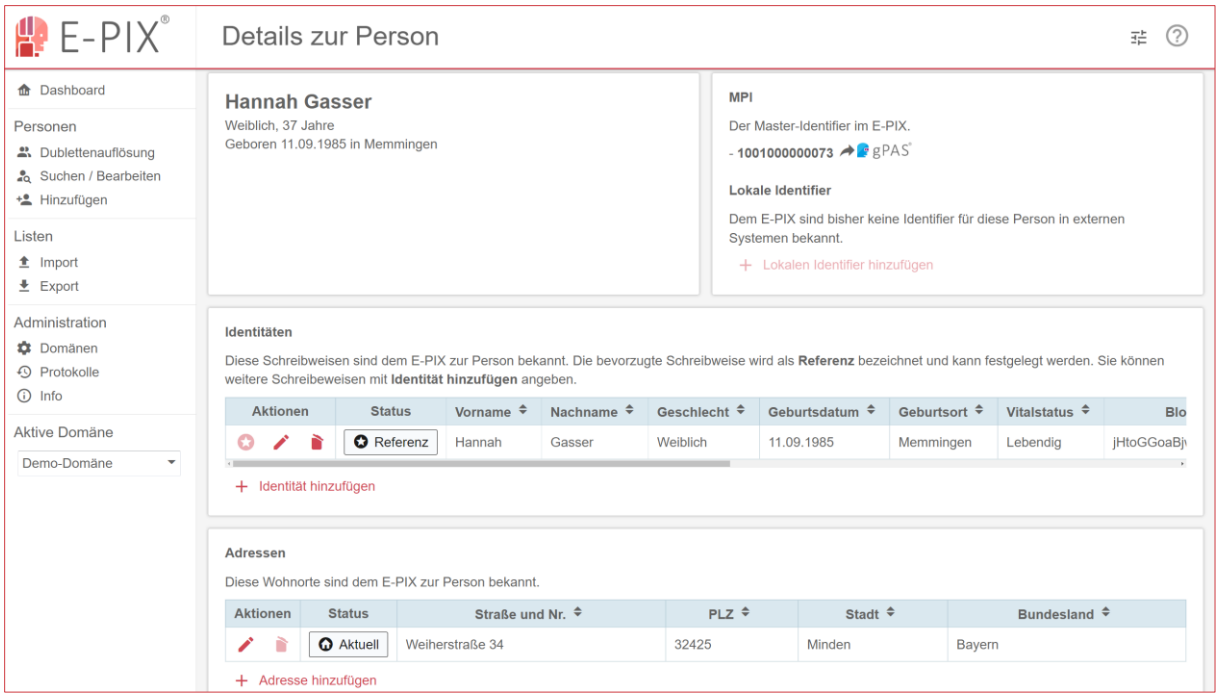


Abbildung 7-4: Detailseite zur Einsicht von den Stammdaten einer Person.

7.5 Bearbeiten und Löschen von Personendaten

Um beispielsweise fehlerhafte Eingaben zu korrigieren oder fehlende Attribute zu ergänzen, kann es erforderlich sein, die Attribute einer Person zu bearbeiten. Hierzu wird zunächst die Detailseite der betreffenden Person aufgerufen (siehe **Abschnitt 7.4**). Jede Identität einer Person kann entsprechend bearbeitet werden. Zur Gewährleistung der Integrität der Daten sollte ein Grund für die Änderungen angegeben werden. Eine Bearbeitung der Stammdaten bedeutet, dass im E-PIX eine neue Identität mit den geänderten Informationen hinzugefügt wird. Daher wird erneut ein Record Linkage durchgeführt.

ⓘ Was passiert, wenn sich die geänderten Stammdaten zu sehr von den Vorherigen unterscheiden?

In diesem Fall teilt der E-PIX dies mit einer Fehlermeldung mit. Die geänderten Daten werden dann nicht übernommen. Um dennoch die neuen Daten zu hinterlegen, kann die Checkbox *Neue Identität erzwingen* ausgewählt werden. Dann werden die neuen Stammdaten in jedem Fall der Person zugeordnet.

Da lediglich eine neue Identität hinzugefügt wird, müssen die alten bzw. fehlerhaften Stammdaten manuell entfernt werden. Standardmäßig werden diese Identitäten nicht gelöscht, da beispielsweise in externen Systemen diese Informationen noch hinterlegt sein könnten und dadurch die Person auch über die zwischenzeitlich geänderten Stammdaten noch im E-PIX auffindbar sein soll. Das Löschen einer Identität ist unwiederbringlich und sorgt dafür, dass jegliche Verweise und Informationen im E-

PIX hierzu gelöscht werden. In **Abbildung 7-5** ist die Oberfläche zum Bearbeiten einer Person abgebildet.

Sind bei einer Person mehrere Identitäten hinterlegt, kann die gewünschte Identität als Referenzidentität ausgewählt werden. Dies kann erforderlich sein, wenn alle Ausprägungen im E-PIX hinterlegt sein sollen, jedoch die korrekte Ausprägung von der gesetzten Referenz-/Hauptidentität abweicht.

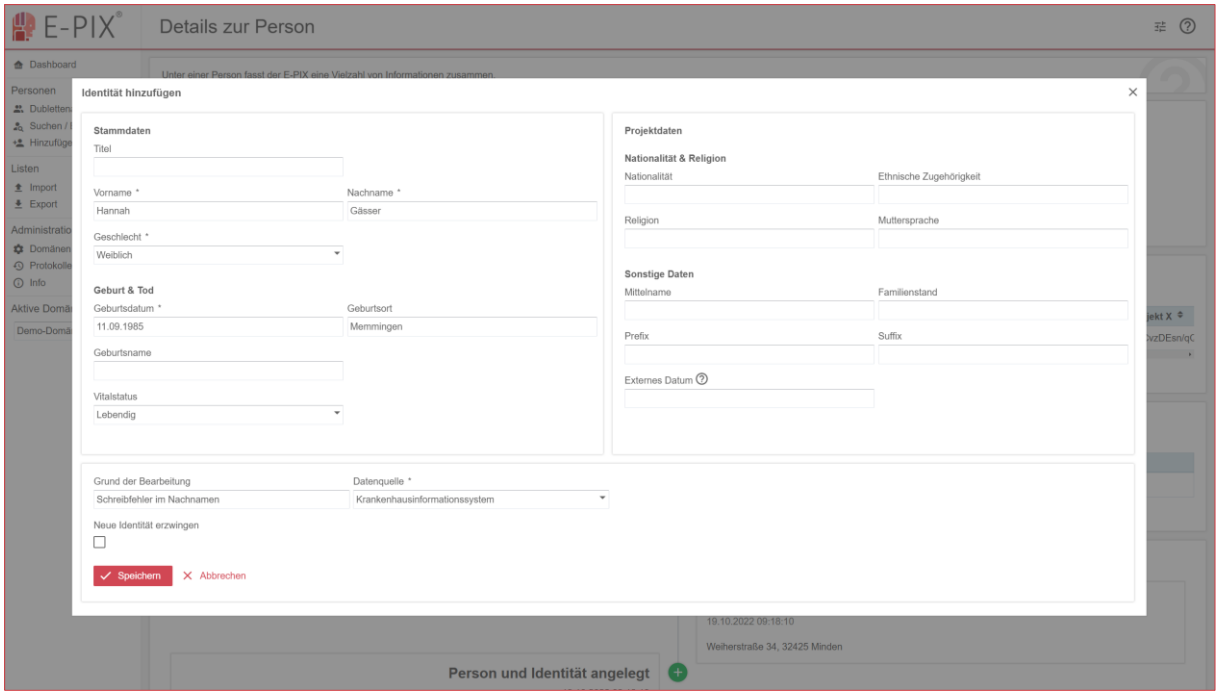


Abbildung 7-5: Oberfläche zum Bearbeiten der Stammdaten einer Person.

Zu jeder Person können beliebig viele Adressen verwaltet werden. Dabei kann ein Eintrag als aktuelle Adresse markiert werden. Beim Hinzufügen neuer Einträge wird stets die neuste Adresse als aktuell markiert. Unabhängig davon kann zu jeder Adresse ein Ein- und Auszugsdatum angegeben werden. Vorhandene Einträge können dupliziert und direkt bearbeitet werden. Vorhandene Einträge können entfernt werden.

7.6 Auflösen möglicher Synonymfehler

Zum Auflösen möglicher Synonymfehler, kann unter dem Menüpunkt *Dublettenauflösung* die Liste möglicher Dubletten eingesehen werden. Um einen möglichen Match aufzulösen, wird ein Eintrag aus der Liste angewählt. Beide Personendatensätze werden tabellarisch gegenübergestellt und Unterschiede bei den jeweiligen Attributen farbig hervorgehoben (siehe **Abbildung 7-6**). So ist eine Entscheidung, ob es sich um ein und dieselbe Person oder zwei unterschiedliche Personen handelt komfortabel möglich. Handelt es sich um zwei Datensätze zu einer natürlichen Person, wird mit der Schaltfläche *Zusammenführen zur Person 1/2* der jeweilige Datensatz als korrekte Ausprägung ausgewählt. Der jeweils andere Datensatz wird der Person als Nebenidentitäten zugeordnet (dabei bleiben alle etwaigen Nebenidentitäten der beiden Personen erhalten). Wenn beide Datensätze zwei unterschiedlichen Personen zugehörig sind, bzw. keine Dublette darstellen, wird über die Schaltfläche

Trennen ein Ausschluss als potentielle Dublette angegeben. Die Personen bleiben dabei getrennt und die Einträge werden aus der Dublettenauflösung entfernt. Für jede Dublettenauflösung kann ein entsprechender Kommentar hinterlegt werden, sodass auch später nachvollzogen werden kann, anhand wessen Kriterien die Entscheidung getroffen wurde. Projektspezifische Begründungen können in der Domänenkonfiguration (siehe **Abschnitt 15.4.9**) definiert werden und sind dann wählbar. Dies reduziert bei häufig auftretenden Fehlern die Schreibarbeit.

Sollte eine Dublettenauflösung nicht sofort möglich sein, weil beispielsweise zunächst weitere Informationen eingeholt werden müssen, kann die Auflösung zurückgestellt werden (Schaltfläche *Zurückstellen*). Damit wird der Eintrag aus der Liste der offenen Dubletten entfernt. Zurückgestellte Dubletten können über die Schaltfläche *Zurückgestellte anzeigen* eingesehen werden. Beide Listen werden gleichermaßen bedient. Zurückgestellte Dubletten können bei Bedarf wieder als offene Dubletten (Schaltfläche *Als offen markieren*) angezeigt werden. Beide Listen können zudem als CSV-Datei exportiert werden.

Wenn zwei Identitäten nicht ähnlich genug sind, um automatisch als *Mögliche Dublette* erkannt zu werden, kann händisch ein entsprechender Eintrag angelegt werden. Hierzu kann die Schaltfläche *Manuell eine Dublette hinzufügen* angewählt werden. Dabei können Dubletten zwischen Personen oder Identitäten angegeben werden. Zwischen zwei Personen werden hierzu die entsprechenden MPIs, bei zwei Identitäten werden die jeweiligen IDs angegeben. Danach erfolgt die Auflösung wie zuvor beschrieben.

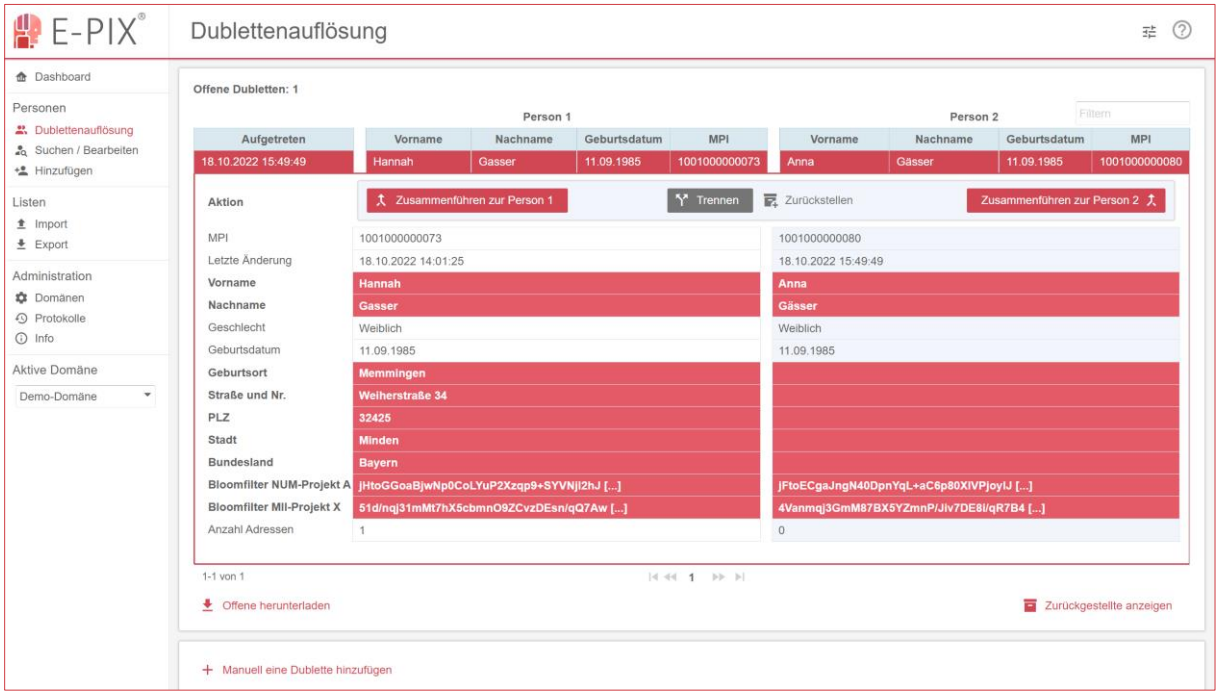


Abbildung 7-6: Gegenüberstellung von Personendaten zum Auflösen einer Dublette.

7.7 Daten exportieren

Die registrierten Personendaten können als CSV-Datei exportiert werden. Hierzu wird unter dem Menüpunkt *Export* der Modus gewählt, anhand dessen die Liste der zu exportierenden Personendaten

bestimmt wird. Personendaten können entweder vollständig oder gefiltert nach einer bestimmten Identifier-Domäne oder anhand bestimmter Stammdaten exportiert werden. Je nach Modus können verschiedene Optionen gewählt werden. Die zu exportierenden Personendaten werden nach der Anwahl der Schaltfläche *Suche* in einer Vorschau angezeigt. Dabei können die zu exportierenden Spalten bestimmt werden, indem durch Anwählen des *X* oder *+* die jeweilige Spalte aus- oder einbezogen wird. Außerdem kann die Reihenfolge der Attribute des resultierenden Exports durch verschieben der Spalten beeinflusst werden. Die resultierende CSV-Datei wird mit der Anwahl der Schaltfläche *CSV herunterladen* heruntergeladen. Die Spalten in der resultierenden Datei werden standardmäßig mit einem Semikolon separiert. Daher enthält die Datei in der ersten Zeile ein `sep=;`. Falls für den Import (siehe **Abschnitt 7.8**) andere Separatoren verwendet werden sollen, kann darüber das entsprechende Zeichen angegeben werden. In **Abbildung 7-7** wird die entsprechende Oberfläche exemplarisch dargestellt.

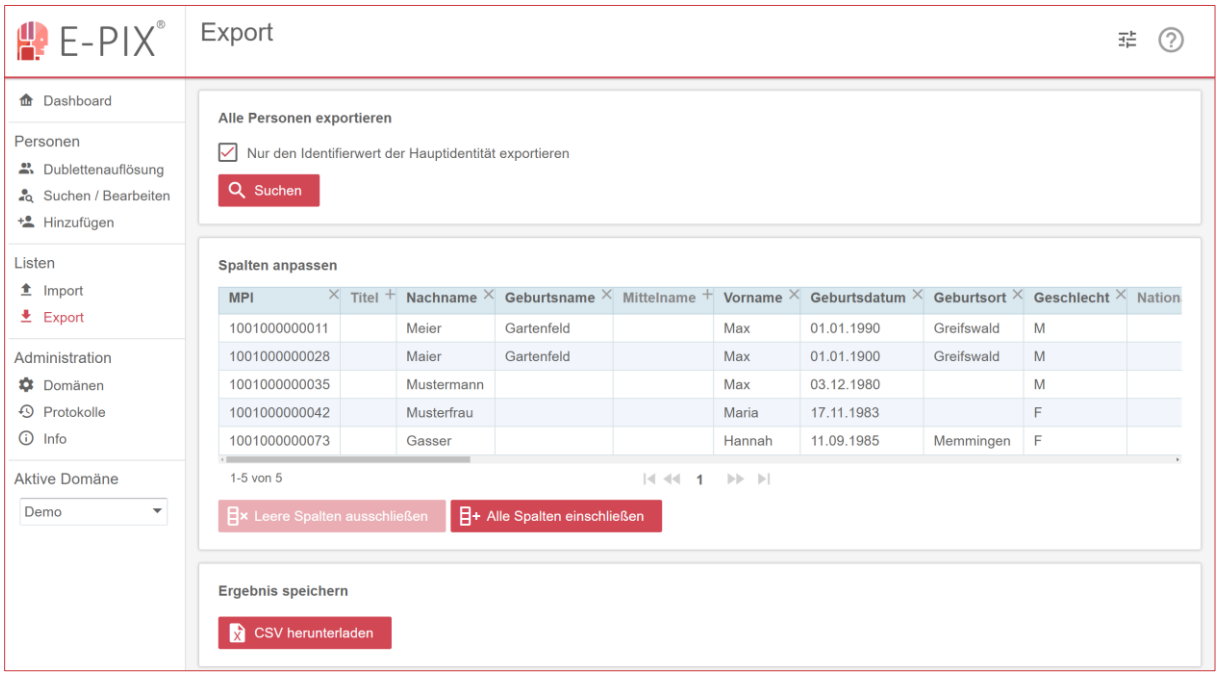


Abbildung 7-7: Oberfläche zum Exportieren von Personendaten.

7.8 Daten importieren

Um Personendaten zu importieren, kann über den Reiter *Import* eine CSV-Datei ausgewählt werden. In **Abbildung 7-8** ist die Oberfläche zum Wählen der CSV-Datei dargestellt.

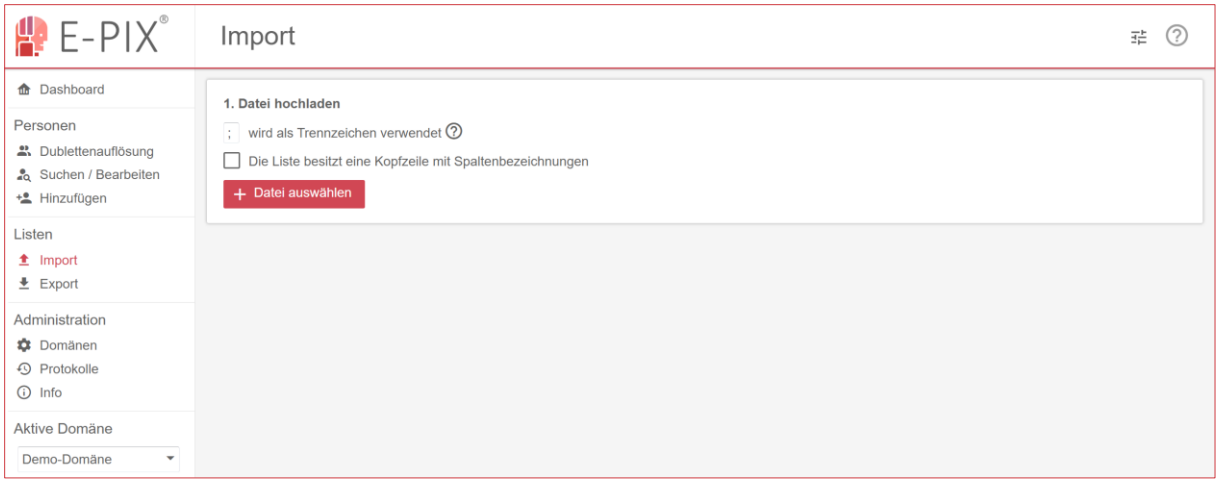


Abbildung 7-8: Oberfläche zum Importieren von Personendaten.

Ist eine Überschrift in der CSV-Datei enthalten, so kann dies mittels der Checkbox *Datei besitzt eine Kopfzeile mit Spaltennamen* mitgeteilt werden. In diesem Fall wird die Kopfzeile nicht mitverarbeitet und führt nicht zu einem Eintrag in den Personendaten. Eine Separierung der Spalten erfolgt standardmäßig mit einem Simikolon. Soll ein anderes Trennzeichen verwendet werden, bspw. ein Semikolon, so kann dies mittels `sep=,` (wenn ein Komma als Separator verwendet werden soll) in der ersten Zeile der CSV-Datei definiert werden[7].

Als Vorschau wird der erste Datensatz aus der Datei dargestellt. Wurden in der CSV-Datei Spaltennamen verwendet, die den Attributnamen der E-PIX-Datenbank entsprechen, erfolgt automatisch eine Zuordnung (z.B. weil die CSV-Datei aus dem E-PIX exportiert wurde (siehe **Abschnitt 7.7**)). Sollen die Spalten anderen Attributen zugeordnet werden oder wurden keine Spaltennamen vorgegeben, so kann über das Auswahlmenü jeder Spalte ein beliebiges Attribut zugewiesen werden. Welche Spalten importiert werden sollen, kann über die Checkboxen bestimmt werden. Einträge mit dem Wert *null* zeigen an, dass es sich um einen Eintrag mit einem leeren Feld handelt. Nach dem Import sind diese Felder entsprechend leer. In **Abbildung 7-9** ist die entsprechende Oberfläche dargestellt.

[7] Dieser Eintrag wird beim Import nicht als Zeile eingelesen und beeinflusst nicht eine etwaig vorhandene Kopfzeile.

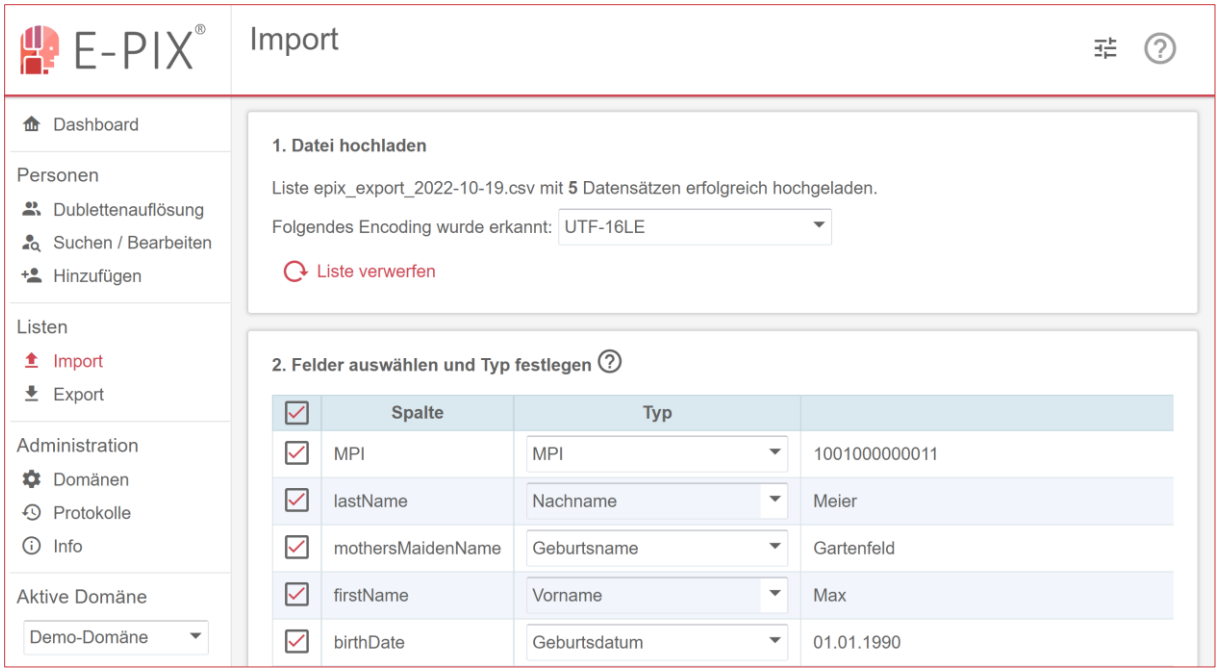


Abbildung 7-9: Oberfläche mit Vorschau der ersten eingelesenen Zeilen.

Für den Import können weitere Optionen festgelegt werden:

- **Datenquelle:** Quelle (siehe **Abschnitt 7.1.1**) der zu importierenden Daten.
- **Kennzeichnung von Änderungen bei einem perfekten Match:** Bei einem *Perfekt Match* bei denen Nicht-Matching-Felder geändert werden, werden diese Datensätze gesondert gekennzeichnet.
- **Vorschau ohne Daten zu speichern:** Der Datenbestand wird nicht verändert. Es wird lediglich das resultierende Ergebnis bei einem Import angegeben.
- **Schreibschutz für Nicht-Matching-Felder:** Bei einem Perfekt Match werden Nicht-Matching-Felder nicht aktualisiert.
- **Schutz beim Import mit MPI vor ungültigen Updates:** Der E-PIX prüft, ob bei identischen MPIs die Stammdaten von Bestandsdaten und zu importierenden Personendaten übereinstimmen und ähnlich genug sind. Wenn keine hinreichende Ähnlichkeit erzielt wird, werden die Daten nicht importiert, bzw. der Person zugeordnet. Diese Option ist standardmäßig aktiviert und kann bei Bedarf deaktiviert werden. Dann werden Identitäten mit geringer Ähnlichkeit einer Person zugeordnet, sofern die MPI übereinstimmt.

ⓘ Was passiert, wenn Stammdaten aus einer Domäne exportiert werden und in eine andere Domäne importiert werden?

Dies ist möglich. Dabei ist zu beachten, dass die Personendaten nur innerhalb einer Domäne eindeutig sind. Das heißt die Personendaten werden nicht domänenübergreifend abgeglichen und entsprechend in jeder Domäne gespeichert. Jedoch müssen die MPIs im E-PIX stets eindeutig sein. Demnach muss beim Import darauf geachtet werden, dass etwaig exportierte MPIs nicht importiert werden. Der E-PIX weißt entsprechend darauf hin, sofern MPIs aus anderen Domänen importiert werden. Der E-PIX vergibt neue MPIs, sofern keine MPIs importiert werden.

7.9 Einsehen von Protokollen

Um nachzuvollziehen, welche Ereignisse eingetreten sind, kann ein Protokoll unter dem Menüpunkt *Protokolle* eingesehen werden. Es stellt dar, welcher Match-Typ durch das Record Linkage für die übertragenden Personendaten errechnet wurde (*Match*, *Möglicher Match*, *Perfekter Match*). Es gibt zudem Aufschluss darüber, ob Personendaten aktualisiert oder Personen neu angelegt oder Identitäten an bestehende Personen angefügt (Nebenidentitäten) wurden. In **Abbildung 7-10** ist eine exemplarische Auflistung dargestellt.

Das angezeigte Protokoll kann anhand der Ereignisse bzw. Events gefiltert werden. Hierzu werden in der Spalte *Ereignis* über eine Auswahlliste die darzustellenden Ereignisse des Record Linkages angewählt. Zudem können die Zeilen nach einer bestimmten Zeichenkette durchsucht werden. Hierfür steht ein Suchfeld zur Verfügung. Dabei werden nur jene Zeilen aufgelistet, welche die entsprechende Zeichenkette in zumindest einer beliebigen Spalte aufweisen.

Das dargestellte Protokoll kann über die Schaltfläche *CSV herunterladen* heruntergeladen werden.

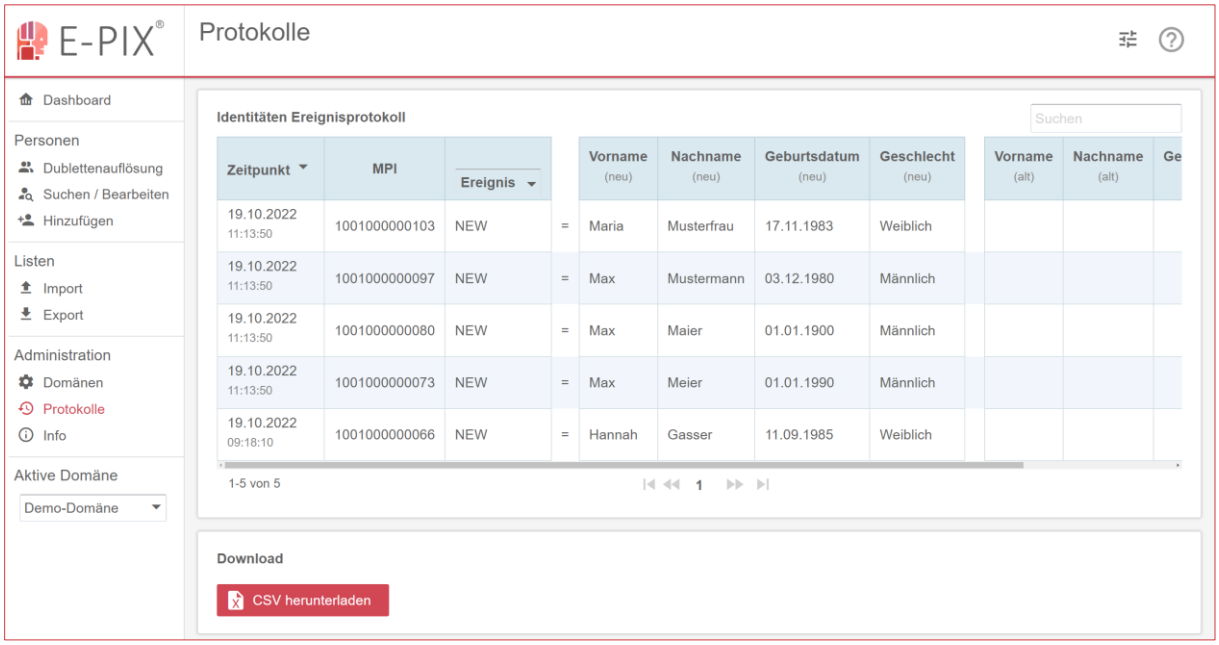


Abbildung 7-10: Oberfläche zum Einsehen des Protokolls.

7.10 Einsehen von Statistiken mittels des Dashboards

Unter dem Menüpunkt *Dashboard* können domänenspezifische und -übergreifende Statistiken eingesehen werden. Hierbei werden diverse Werte wie vorhandene *Mögliche Matches*, registrierte Personen, vorhandene Identitäten, aufgelöste Dubletten (separat aufgeführt als zusammengeführte und getrennte Personen), usw. gelistet und grafisch aufbereitet dargestellt.

Die Statistik kann als CSV über die jeweiligen Schaltflächen heruntergeladen werden. In **Abbildung 7-11** ist exemplarisch eine Statistik dargestellt.

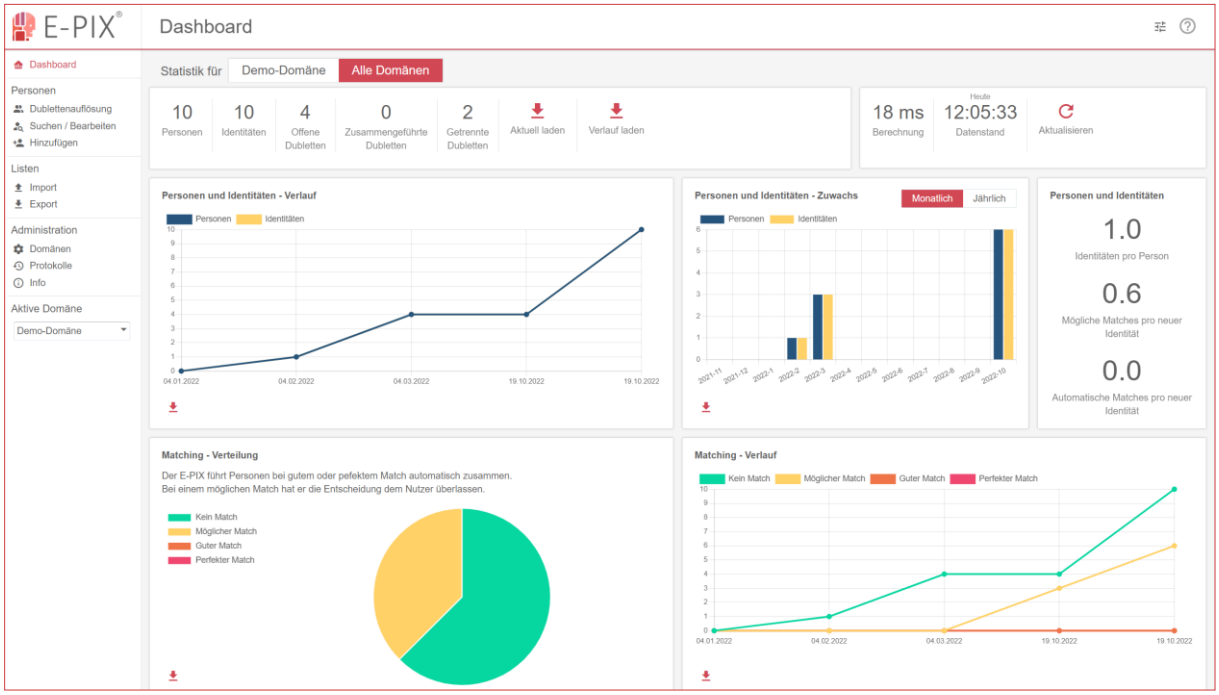


Abbildung 7-11: Dashboard zum Einsehen der Statistiken.

Die gezeigten Statistiken werden asynchron, also nicht automatisch und nicht in Echtzeit, generiert. Die Aktualisierung kann jederzeit manuell über die Schaltfläche *Aktualisieren* angestoßen werden. Die dabei generierten Daten werden durch den E-PIX erzeugt und in der Datenbank dokumentiert. Das Dashboard ersetzt damit die bis E-PIX 2.12.x genutzten Kennzahlenprozeduren innerhalb der E-PIX-Datenbank.

ⓘ Unterstützung bei regelmäßiger Community-Kennzahlenerhebung.

Das Dashboard liefert einen schnellen Überblick über Zahlen zu Personen und Identitäten. Diese können als CSV-Datei exportiert und der Unabhängigen Treuhandstelle Greifswald per E-Mail übermittelt werden. Das unterstützt bei statistischen Auswertungen über die Gesamtzahl von Personen und Identitäten in der Community. Vielen Dank fürs Mitmachen!

8 Logging

⚠ **Hinweis:** Details für die Anpassung der Logging-Konfiguration entnehmen Sie bitte der beigelegten Beschreibung docker-compose/README.md (Abschnitt Logging).

9 Versand von Notifications

Wie in der Architekturgrafik zu sehen (siehe **Abbildung 6-1**), ist der E-PIX seit Version 2.13.0 in der Lage Benachrichtigungen an externe Systeme zu versenden. Dies kann per `http`, `MQTT` oder `EJB` erfolgen. Die Versandmitteilungen werden in einer separaten Notification-Datenbank dokumentiert.

⚠ **Hinweis:** Details zum Umfang der Notification-Schnittstelle, zur Einrichtung, sowie weitere Erläuterungen sind separat unter https://www.ths-greifswald.de/ttp-tools/notifications dokumentiert.

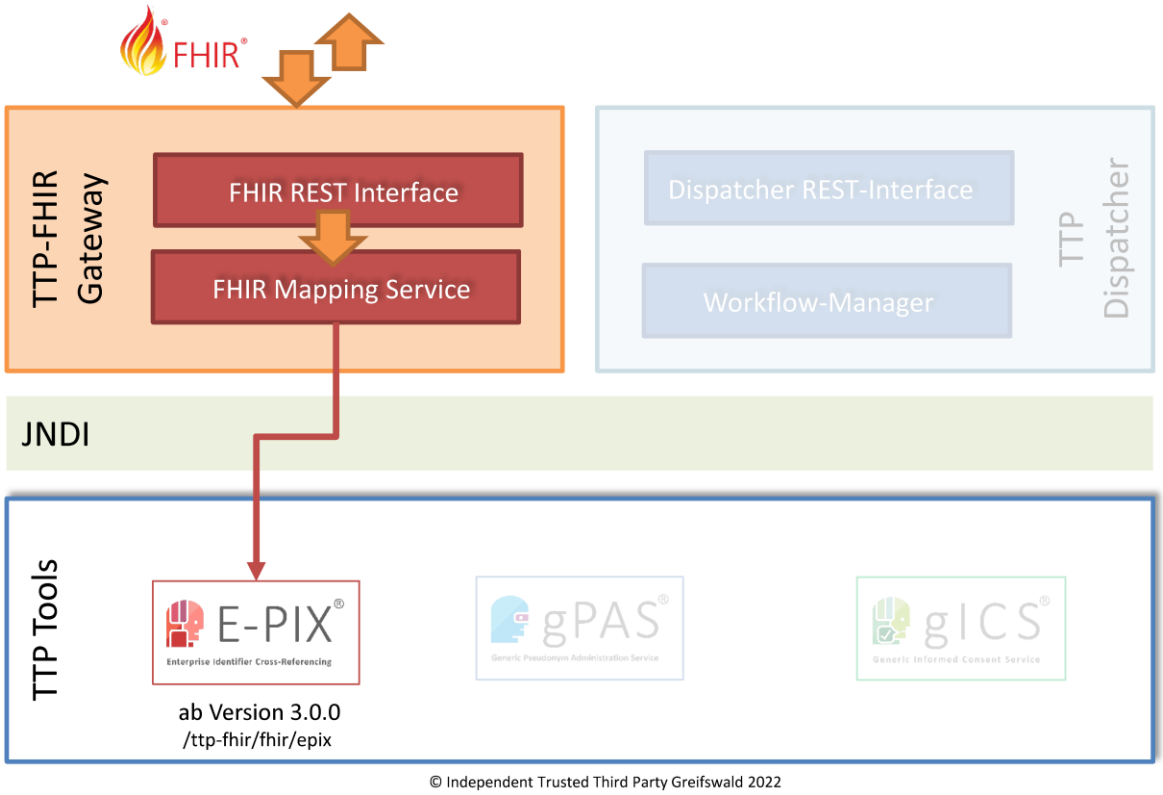
10 FHIR-Unterstützung für E-PIX per TTP-FHIR Gateway

„Fast Healthcare Interoperability Resources (kurz: FHIR®) ist ein von HL7 erarbeiteter Standard. Dieser unterstützt den Datenaustausch zwischen Softwaresystemen im Gesundheitswesen. FHIR beschreibt Datenformate und Elemente als sogenannte „Ressourcen“ und bietet eine Schnittstelle an, um diese auszutauschen“[8].

Um sowohl bestehende Anwenderprojekte als auch künftige Nutzer bei der Umsetzung FHIR-orientierter Infrastrukturen und Prozesse zu unterstützen, wird ab E-PIX-Version 3.0 ein zusätzliches Treuhandstellen-FHIR-Gateway (kurz: TTP-FHIR Gateway) als Mittler von FHIR-spezifischen Infrastrukturkomponenten und E-PIX bereitgestellt.

[8] https://de.wikipedia.org/wiki/Fast_Healthcare_Interoperability_Resources, Zugriff am 22.01.2021

⚠ **Hinweis:** Da der E-PIX als datenhaltendes System sämtliche identifizierende Daten und Pseudonyme erster Stufe (MPI) verwaltet, ist der E-PIX auch für die Generierung und Verwaltung der erforderlichen FHIR-UUIDs verantwortlich.



Für ausgewählte Funktionalitäten zum Anlegen von Personendaten in FHIR wurden nachfolgende Funktionen umgesetzt und sind nach erfolgreichem Deployment des TTP-FHIR Gateways direkt per REST nutzbar.

Der aktuelle Funktionsumfang (FHIR-Operations) des TTP-FHIR Gateway umfasst:

- Anlegen von Personendaten
- Aktualisieren von Personendaten

Darüber hinaus gibt es eine Vielzahl von Suchfunktionen. Weitere Funktionalitäten werden sukzessive implementiert und bereitgestellt. Der zugehörige Implementation Guide mit konkreten Beispielen ist zu finden unter https://www.ths-greifswald.de/e-pix/fhir.

⚠ **Hinweis:** Die Profilierung der erforderlichen Profile, Codesysteme und Operations erfolgte in Zusammenarbeit mit der Fa. gefyra[9].

[9] https://www.gefyra.de/, Zugriff am 2021-06-08

11 Authentifizierung und Autorisierung

Die bereitgestellte E-PIX®-Version bietet unterschiedliche Umsetzungsoptionen der Authentifizierung und Autorisierung sowohl in der Docker- als auch in der Docker-Compose-Variante.

In der Standard-Ausgabe vom E-PIX® ist keine Authentifizierung notwendig. Soll der E-PIX® nur für bestimmte Nutzergruppen (Admin-Nutzer, Standard-Nutzer) zugänglich gemacht werden (vgl. **Tabelle 11-1**) oder das Anlegen von neuen Domänen beschränkt werden, stehen dafür zwei Authentifizierungsverfahren bereit. gRAS und KeyCloak, wobei es für KeyCloak zwei verschiedene Varianten gibt. *Die Verwendung von KeyCloak wird empfohlen*.

⚠ **Hinweis:** Die Auswahl der einzelnen Varianten erfolgt in der Docker-Compose Version innerhalb der docker-compose.yml. Details für die notwendige Anpassung der Docker-Konfiguration können der beigelegten Beschreibungen https://www.ths-greifswald.de/ttp-tools/keycloak sowie docker-compose/README.md.

⚠ **Hinweis:** Mit dem Herbstrelease 2022 können nun **alle THS-Schnittstellen (WEB-Oberfläche, FHIR-Gateway und SOAP-Webservices)** je Endpunkt und somit je Werkzeug (E-PIX, gICS, gPAS) mit KeyCloak-basierter (und damit OIDC-konformer) Authentifizierung abgesichert werden.

Die Konfiguration der Authentifizierung erfolgt in der Docker-Compose Version innerhalb der `ttp_epix.env`. Eine detailliierte Beschreibung ist unter https://www.ths-greifswald.de/ttp-tools/keycloak verfügbar.

11.1 Übersicht Nutzerrollen und Rechte

Tabelle 11-1: Nutzer der Gruppe Admin **und** User **haben unterschiedliche Zugriffsrechte in der Web-Oberfläche.**

Bereich/Seite	Zugang ohne Login	Zugang mit User-Rechten	Zugang mit Admin-Rechten
Info	×	×	×
Dashboard		×	×
Administration: Domänen			×
Administration: Protokolle		×	×
Administration: Statistik		×	×
Personen: Dublettenauflösung		×	×
Personen: Suche / Bearbeiten		×	×
Personen: Hinzufügen		×	×

Listen: Import	✕
Listen: Export	✕
Administration: Domänen	✕

12 Verwendung von KeyCloak

⚠ **Hinweis:** Details zur Vorbereitung des KeyCloak-Servers sind unter https://www.ths-greifswald.de/ttp-tools/keycloak beschrieben.

Die Client-seitige KeyCloak-Konfiguration kann sowohl per Config-Datei als auch per Environment-Variablen bei Start des Docker-Compose erfolgen. Details sind in **docker-compose/README.md** beschrieben.

Neben der Absicherung der Weboberfläche gibt es die Möglichkeit, die SOAP-Schnittstelle per KeyCloak abzusichern. Hierfür wird ähnlich wie bei der Weboberfläche in Zugriffsrechte für Admin und User unterschieden.

12.1 Verwendung von gRAS

⚠ **Hinweis:** Details zur Administration und Nutzung der gRAS-Authentifizierung sind unter folgendem Link https://www.ths-greifswald.de/ttp-tools/gras am Beispiel von gPAS® dokumentiert.

13 Empfehlungen zur Absicherung des Anwendungsservers

Der Zugriff auf relevante Anwendungs- und Datenbankserver des E-PIX sollte nur für autorisiertes Personal und über autorisierte Endgeräte möglich sein. Wir empfehlen die Umsetzung nachfolgender IT-Sicherheitsmaßnahmen:

- Betrieb der relevanten Server in separaten Netzwerkzonen (getrennt von Forschungs- und Versorgungsnetz)
- Verwendung von Firewalls und IP-Filtern
- Verwendung von KeyCloak
- Zugangsbeschränkung auf URL-Ebene mit Basic Authentication (z.B. mit NGINX oder Apache)

14 Nutzung der SOAP-Schnittstelle

Neben der grafischen Benutzerschnittstelle, steht eine maschinenverständliche Web-Schnittstelle zur Verfügung. Diese kann mit dem SOAP-Protokoll angesprochen werden. Beim laufenden Dienst werden

je nach Zweck die dazu vorhandenen Definitionen der SOAP-Schnittstellen mit dem folgenden Pfaden abgerufen (die URLs müssen entsprechend angepasst werden).

Personenverwaltung (inkl. Record Linkage) und ID Administration:

```
http://example.org:8080/epix/epixService?wsdl
```

Konfiguration und Domänenmanagement:

```
http://example.org:8080/epix/epixManagementService?wsdl
```

Versenden von Notifications:

```
http://example.org:8080/epix/epixServiceWithNotification?wsdl
```

Die dazugehörige Entwicklerdokumentation ist unter der folgenden URL zu finden.

```
https://www.ths-greifswald.de/epix/doc
```

14.1 Registrierung von Personen

Im Folgenden wird exemplarisch die Registrierung einer Person vorgestellt. Mit der Funktion `requestMPI` wird die entsprechende Person registriert, sofern diese noch nicht in der jeweiligen Domäne enthalten ist. Wenn die Person bereits registriert wurde, bzw. Personendaten mit einer sehr hohen Übersteinstimmung vorhanden sind, so wird keine neue Person angelegt, sondern eine neue Identität der entsprechenden Person zugeordnet (vgl. Nebenidentitäten in **Abschnitt 4**). Der Abgleich findet mittels Record Linkage statt. Das Ergebnis dieses Vorgangs (Match-Typ, vgl. **Tabelle 7-1**) wird in der Antwort auf die Anfrage geliefert. Zudem wird ein MPI (vgl. **Abschnitt 2**) geliefert, der für die angegebene Domäne eindeutig ist. Eine exemplarische Anfrage ist in **Abbildung 14-1** abgebildet.

```
<soapenv:Envelope xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
     xmlns:ser="http://service.epix.ttp.icmvc.emau.org/">
   <soapenv:Header/>
   <soapenv:Body>
      <ser:requestMPI>
         <domainName>Dummy</domainName>
         <identity>
            <birthDate>1985-05-21</birthDate>
            <firstName>Maximilian</firstName>
            <gender>M</gender>
            <lastName>Mustermann</lastName>
            <contacts>
               <city>Musterstadt</city>
               <street>Musterstraße 11</street>
               <zipCode>12345</zipCode>
            </contacts>
         </identity>
         <sourceName>dummy_safe_source</sourceName>
      </ser:requestMPI>
   </soapenv:Body>
</soapenv:Envelope>
```

Abbildung 14-1: Exemplarische Anfrage zur Registrierung einer Person.

Die Antwort auf die eben gezeigte Anfrage ist in **Abbildung 14-2** dargestellt.

```
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
   <soap:Body>
      <ns2:requestMPIResponse xmlns:ns2="http://service.epix.ttp.icmvc.emau.org/">
         <return>
            <matchStatus>NO_MATCH</matchStatus>
            <person>
               <mpiId>
                  [...]
                  <value>1001000000059</value>
               </mpiId>
               [...]
            </person>
         </return>
      </ns2:requestMPIResponse>
   </soap:Body>
</soap:Envelope>
```

Abbildung 14-2: Gekürzte Antwort auf die Anfrage zur Registrierung einer Person.

Mit der Funktion `requestMPIBatch` können mehrere Personen innerhalb einer Anfrage registriert werden.

Wenn eine Person angelegt wurde und nachträglich Attribute verändert werden sollen, erfolgt dies mit der Funktion `updatePerson`.

ⓘ Was bedeutet der `matchCode` MULTIPLE-MATCH?

Der übergebene Datensatz führt würde bei mehreren Personen zu einem *automatischen Match* (siehe **Tabelle 7-1**) führen. Deshalb werden in diesem Fall die Datensätze nicht zusammengeführt, sondern eine neue Person angelegt und eine zugehörige Liste mit *möglichen Matches* (siehe **Tabelle 7-1**) angelegt.

ⓘ Welche MPI-ID wird bei einem *possible Match/möglichen Match* zurückgegeben?

Bei einem *möglichen Match* (siehe **Tabelle 7-1**) wird zunächst eine neue Person angelegt und eine neue MPI-ID erzeugt. Die *möglichen Matches* können im Nachgang aufgelöst werden. Zurückgeliefert wird die neue MPI-ID der neu angelegten Person.

14.2 Personen per MPI suchen

Innerhalb einer Domäne kann mittels des eindeutigen MPI eine Person gesucht werden. Hierzu steht die Funktion `getPersonByMPI` bereit. In der entsprechenden Anfrage muss der Domänenname und der MPI der gesuchten Person angegeben werden. In **Abbildung 14-3** ist eine exemplarische Anfrage zum Suchen einer Person mittels des MPIs dargestellt.

```
<soapenv:Envelope xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
     xmlns:ser="http://service.epix.ttp.icmvc.emau.org/">
   <soapenv:Header/>
   <soapenv:Body>
      <ser:getPersonByMPI>
         <domainName>Dummy</domainName>
         <mpiId>1001000000059</mpiId>
      </ser:getPersonByMPI>
   </soapenv:Body>
</soapenv:Envelope>
```

Abbildung 14-3: Exemplarische Anfrage zum Suchen einer Person mittels des dazugehörigen MPIs.

In **Abbildung 14-4** ist ein Auszug aus der entsprechenden Antwort dargestellt. In der Antwort sind alle Personendaten enthalten.

```
<soap:Envelope xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
   <soap:Body>
      <ns2:getPersonByMPIResponse xmlns:ns2="http://service.epix.ttp.icmvc.emau.org/">
         <return>
            <mpiId>
               [...]
               <value>1001000000059</value>
            </mpiId>
            [...]
         </return>
      </ns2:getPersonByMPIResponse>
   </soap:Body>
</soap:Envelope>
```

Abbildung 14-4: Antwort auf die Anfrage zum Suchen einer Person mittels des MPIs.

Mit der Funktion `getPersonByLocalIdentifier` können statt mittels MPI mit einem Lokalem Identifier die dazugehörigen Personendaten abgerufen werden.

14.3 Alle Personendaten zu einer Domain

Mit der Funktion `getPersonsForDomain` können alle Personendaten aller Personen einer Domäne abgerufen werden. Hierzu muss in der entsprechenden Anfrage die jeweilige Domäne angegeben werden. In der Antwort sind alle Personendaten aufgelistet. In **Abbildung 14-5** ist exemplarisch eine Anfrage dargestellt.

```
<soapenv:Envelope xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
      xmlns:ser="http://service.epix.ttp.icmvc.emau.org/">
   <soapenv:Header/>
   <soapenv:Body>
      <ser:getPersonsForDomain>
         <domainName>Dummy</domainName>
      </ser:getPersonsForDomain>
   </soapenv:Body>
</soapenv:Envelope>
```

Abbildung 14-5: Anfrage um alle Personendaten einer Domain abzurufen.

Analog dazu können mit der Funktion `getIdentitiesForDomain` alle Identitäten aus einer Domäne abgerufen werden.

15 Konfiguration von E-PIX Domänen

Jedes Projekt und jedes Forschungsvorhaben haben unterschiedliche Anforderungen bei der technischen Umsetzung zu berücksichtigen. Register, wie das Klinische Krebsregister MV (KKR-MV), verzeichnen alle Krebspatienten aus Mecklenburg-Vorpommern. Hier ist eine besonders hohe Genauigkeit bei der Zusammenführung von Informationen (bei bislang mehr als 255.000 Personen) aus den beteiligten Registerstellen und bei der Identifikation der Personen erforderlich. Jede Abweichung in den demografischen Informationen, sei es nur ein Zeichen, soll dem Treuhandstellenpersonal signalisiert werden und muss einer genauen Prüfung unterzogen werden.

In der NAKO Gesundheitsstudie werden die demografischen Daten der potentiellen Studienteilnehmer von den Meldeämtern abgerufen. Da hier von einer gewissen Grundqualität der Daten auszugehen ist, sind die Schwellwerte deutlich höher als im KKR-MV gewählt. Dies hat zur Folge, dass bei mehr als 2

Mio. eingeschlossenen Personen die nötige manuelle Nacharbeit zum Auflösen potentieller Matches, bei gleichzeitiger Gewährleistung der Qualität, auf ein Mindestmaß reduziert werden konnte.

Beide Beispiele lassen sich problemlos über entsprechende Schwellwerte und Parameter mit Hilfe der E-PIX Konfiguration abbilden.

Grundlage der Erkennung der Personen, ist der Matching-Prozess des E-PIX. Das beabsichtigte Verhalten (welche Felder sollen wie abgeglichen werden) und die nötige Genauigkeit (wann soll der E-PIX entscheiden und wann sollen potentielle Matches signalisiert werden) kann über einer Vielzahl von Schwellwerten und Parametern konfiguriert werden.

Je niedriger die Schwellwerte für potentielle Matches gewählt werden desto mehr Matching-Paare von Personen werden signalisiert und umso mehr manuelle Kontrolle dieser möglichen Matches durch das Treuhandstellenpersonal ist erforderlich.

Die Konfiguration des E-PIX erfolgt je Domäne. Um die Vielzahl der Anpassungsmöglichkeiten zu verstehen, werden nachfolgende grundlegend die Matching-Mechanismen und die möglichen Konfigurationsoptionen beschrieben.

⚠ **Hinweis:** Die Konfiguration des E-PIX sollte stets vor produktivem Beginn des Vorhabens erfolgen. Der E-PIX entscheidet über den Matching-Zustand einer Person auf Basis der bereits vorhandenen Daten und der aktuellen Konfiguration. Aktualisiert man die Konfiguration bzgl. des Matchings oder der Aufbereitung der Eingabedaten, obwohl bereits Daten in der Datenbank vorhanden sind, müssen diese erneut eingespielt werden (idealerweise in ein leeres System), um die Korrektheit der Matching-Bewertung gemäß der neuen Konfiguration gewährleisten zu können.

⚠ **Hinweis:** Standardkonfigurationen (mit und ohne Bloomfilter) werden beim E-PIX mitgeliefert und können als Grundlage für Änderungen oder Erweiterungen verwendet werden. Zu finden sind diese im Verzeichnis `/examples` als .xml-Dateien. Zudem befindet sich eine Demo-Datenbank (`.sql`) mit exemplarischen Daten.

15.1 Hintergrund

Für die Registrierung eines Personendatensatzes, kann der Matching-Prozess in mehrere Teilschritte gegliedert werden. Hierzu wird zunächst eine Vorselektion der Personendatensätze vorgenommen (Blocking), sodass eine verringerte Anzahl von Datensätzen für ein unschärferes Matching (siehe **Abschnitt 15.1.1**) abgeglichen werden muss. Die Personendatensätze, die dabei eine hinreichende Ähnlichkeit aufweisen, werden mithilfe eines genaueren Abgleichs unter Zuhilfenahme von höheren Schwellwerten und weiteren Feldern vergleichen (siehe **Abschnitt 15.1.2**). Je nach Ähnlichkeit können die verbleibenden Datensätze als Dublette, keine Dublette oder mögliche Dublette klassifiziert werden. In **Abbildung 15-1** ist der Prozess der Registrierung stark vereinfacht dargestellt.

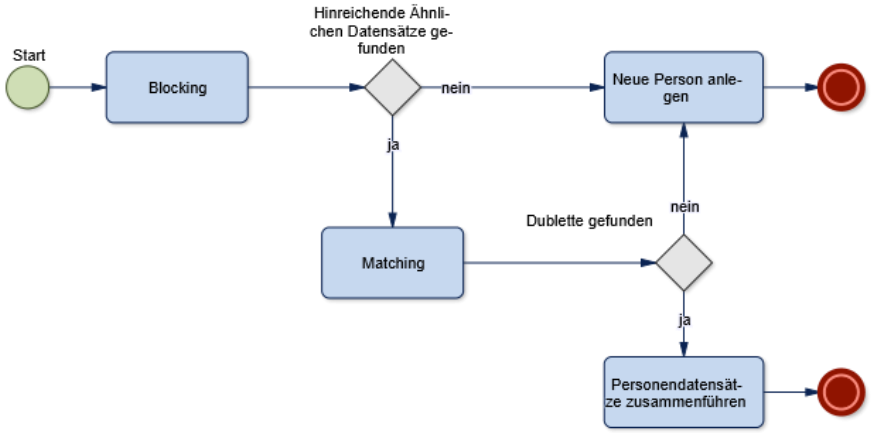


Abbildung 15-1: Vereinfachter Ablauf des Matching-Prozesses.

15.1.1 Blocking

Grundsätzlich dient das Blocking einer ersten unscharfen Selektierung potentieller Duplikate. Im E-PIX ist dieser Vorgang frei konfigurierbar, indem eine reduzierte Teilmenge von Attributen verwendet wird, um einen ersten Abgleich durchzuführen. Dabei wird eine Ähnlichkeit zwischen jeweils zwei Datensätzen (dem zu registrierenden und einem bereits registrierten Datensatz) ermittelt. Die Attribute, welche hierzu abgeglichen werden sollen, können gewählt werden und betreffend des Schwellwerts konfiguriert werden (siehe **Abschnitt 15.4.12**). Das Blocking dient der Steigerung der Performance und verringert insbesondere bei großen Datenbeständen die Dauer eines Abgleichs.

15.1.2 Matching

Wird beim Blocking eine hinreichende Ähnlichkeit mit bestimmten Datensätzen ermittelt, so werden diese in einem weiteren Abgleich genauer verglichen. Hierzu werden weitere Attribute hinzugezogen, welche ebenfalls ähnlich wie beim Blocking konfiguriert (siehe **Abschnitt 15.4.12**) werden können. Mithilfe dieses genaueren Abgleichs kann klassifiziert werden, ob ein Duplikat vorliegt oder nicht[10].

15.2 XML-basierte Konfiguration

Die Konfiguration des E-PIX wird im XML-Format definiert. Über das Web-Frontend des E-PIX kann die Konfiguration von E-PIX-Domänen (Projekte, Studien, Quellsysteme) angezeigt und editiert werden.

[10] Tatsächlich findet eine feinere Unterteilung statt (vgl. **Tabelle 7-1**)

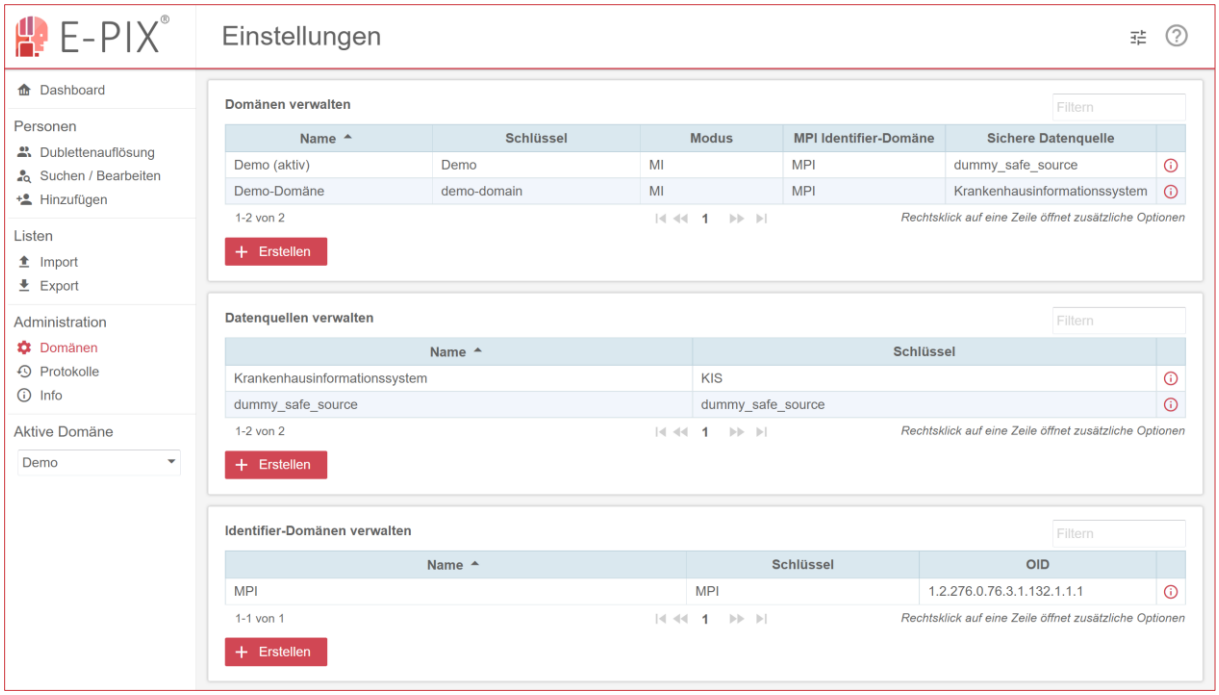


Abbildung 15-2: Das Anzeigen und Editieren der aktuellen Konfiguration einer E-PIX-Domäne ist direkt über das Web-Frontend möglich.

In **Abbildung 15-3** ist die Struktur der Konfiguration illustriert. Es sind alle Elemente aufgelistet, die bei der Domänenkonfiguration verwendet werden können. Die Struktur gibt dabei an, welche Elemente anderen Elementen untergeordnet sind. Eine Erläuterung aller Elemente mit Beispielen und validen Wertebereichen folgt im nächsten Abschnitt.

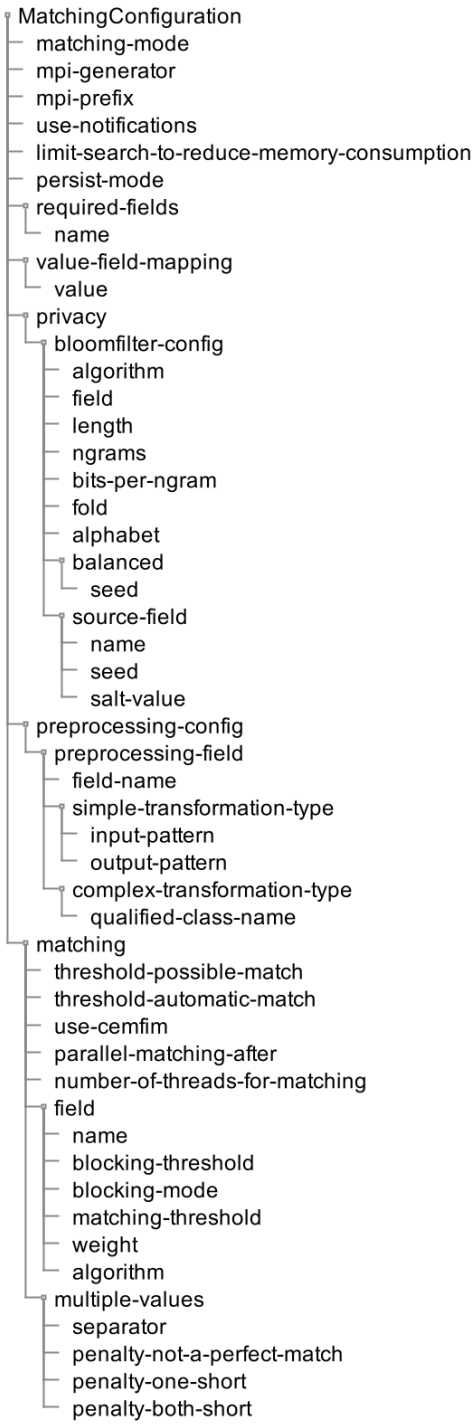


Abbildung 15-3: Alle Elemente, die bei der Konfiguration der Domäne verwendet werden können.

Im E-PIX sind mehrere Felder vorgegeben. Je nach Feld wird standardmäßig eine formale Prüfung von Eingaben durchgeführt. So würde beispielsweise der 31.02. nicht als Geburtsdatum angenommen werden. In

Tabelle 15-1 sind alle vordefinierten Felder aufgelistet. Hierbei ist zu beachten, dass die Felder der Kontaktdaten nicht für das Matching verwendet werden können.

Tabelle 15-1: Alle im E-PIX definierten Felder.

Feldname	Beschreibung	Beispiel
firstName	Vorname	Anna
middleName	Weitere Vornamen	Lea
lastName	Nachname	Schmidt
birthDate	Geburtsdatum Format: JJJJ-MM-TT	1980-03-12
gender	Geschlecht (wird intern auf mittels eines Buchstaben angegeben) *m* für male (männlich), *f* für female (weiblich), *o* für other (sonstige), *u* für Unknown (unbekannt) und *x* für divers	f
externalDate	Freies Feld für ein Datum, welches im E-PIX nur gespeichert, aber nicht weiter prozessiert wird Format: JJJJ-MM-TT	2019-04-30
birthPlace	Geburtsort	Berlin
race	Ethnizität	Kaukasier
religion	Religion	Christentum
mothersMaidenName	Geburtsname	Müller
degree	Abschluss	Mittlerer Schulabschluss
motherTongue	Muttersprache	deutsch
nationality	Nationalität/Staatsangehörigkeit	deutsch
civilStatus	Familienstand	ledig
value1 - value10	Felder dessen Werte je Projekt/Studie individuell belegt werden können. Zusätzlich kann der Feldname für die Weboberfläche mittels eines Labels geändert werden (siehe **Abschnitt 15.4.8**). Die Felder haben in der Datenbank unterschiedliche Längen und sind auf diese limitiert: value1-5 max. 50 Zeichen value6 und 7 max. 255 Zeichen	Todesdatum

	value8 und 9 max. 1.000 Zeichen value10 max. 10.000 Zeichen	
prefix	Präfix (Name), Vorsatzwort	von
suffix	Suffix (Name), Namenszusatz	B. Sc.
city	Wohnort (Kontaktdaten)	Berlin
country	Land (Kontaktdaten)	Deutschland
countryCode	Ländercode (Kontaktdaten)	49
district	Bezirk/Stadtteil (Kontaktdaten)	Spandau
email	E-Mail-Adresse (Kontaktdaten)	anna.schmidt@beispiel.de
externalDate	Freies Feld für ein Datum, welches im E-PIX nur gespeichert, aber nicht weiter prozessiert wird (Kontaktdaten) Format: JJJJ-MM-TT	2019-06-27
municipalityKey	Amtlicher Gemeindeschlüssel (Kontaktdaten)	11000000
phone	Telefonnummer (Kontaktdaten)	030/123 456 789
state	Bundesland (Kontaktdaten)	Berlin
street	Straße (Kontaktdaten)	Spandauer Damm
zipCode	Postleitzahl (Kontaktdaten)	13593
comment	Kommentar	*beliebig*
vitalStatus	Vitalstatus Unterstützte Werte sind: *ALIVE* (lebendig), *DEAD* (verstorben), *UNKNOWN* (unbekannt)	*ALIVE*
dateOfDeath	Sterbedatum	2015-03-20

15.3 Die Standard-Konfiguration

Dem E-PIX ist eine Standard-Konfiguration für Domänen beigelegt. Diese kann ohne Weiteres für viele Projekte vorerst ausreichend sein. Hierbei ist jedoch, wie oben bereits erwähnt zu beachten, dass eine nachträgliche Änderung der Domänen-Konfiguration für eine korrekte Bewertung des Matchings eine komplette Neuregistrierung aller bereits bekannten Datensätze nach sich ziehen muss.

Die Standard-Konfiguration nutzt für das Record Linkage die Felder `firstName` (Vorname), `lastName` (Nachname), `birthDate` (Geburtsdatum) und `gender` (Geschlecht). Die Felder

`firstName` und `lastName` werden für den Abgleich mittels pre-processing (siehe **Abschnitt 15.4.11**) aufbereitet. Für das Blocking werden die Felder `firstName` und `birthDate` verwendet. Für das Feld `firstName` werden zudem Multiple-Values (siehe **Abschnitt 15.4.12.7**) genutzt. Ein Matching findet mithilfe aller vier Felder statt. Für einen Abgleich wird immer die Levenshtein-Distanz verwendet[11]. In **Tabelle 15-2** sind die Felder zur Übersicht dargestellt.

Tabelle 15-2: Verwendete Felder mit Schwellwerten und Wichtung in der Standard-Domänenkonfiguration.

Feldname	Blocking-Schwellwert	Matching-Schwellwert	Wichtung des Felds
firstName	0,4	0,8	8
lastName	*Nicht für Blocking verwendet*	0,8	6
birthDate	*Nicht für Blocking verwendet*	0,75	3
gender	0,6	1,0	9

15.4 Struktur und Inhalt der Konfiguration

Das Element `MatchingConfiguration` ist das Wurzelelement. Alle Elemente sind diesem Element untergeordnet.

15.4.1 matching-mode

Mithilfe des Elements `matching-mode` kann definiert werden, ob ein Record Linkage durchgeführt werden soll, oder nicht. Mit dem Modus `MATCHING_IDENTITIES`, findet ein Record Linkage statt. Mit dem Modus `NO_DECISION` wird kein Record Linkage durchgeführt und Personendaten werden nur übernommen und im E-PIX hinterlegt. Dies kann gewünscht sein, wenn Personendaten z.B. durch ein KAS/KIS übermittelt werden und bereits Identifizierer vergeben wurden und bereits ein Record Linkage durchgeführt wurde. In **Tabelle 15-3** sind die zwei Modi im Detail erläutert.

Tabelle 15-3: Unterstützte Matching-Modes

Wert	Beschreibung
`MATCHING_IDENTITIES`	Bei der Registrierung von Personen wird ein Record Linkage durchgeführt (Verwendung von `addPerson` nicht möglich). Die Konfiguration des Record Linkages wird mit dem Element `matching` angegeben.
`NO_DECISION`	Bei der Registrierung von Personen findet kein Record Linkage statt und die Personendaten werden nur übernommen. Bei jedem

[11] Weitere Vergleichsmöglichkeiten sind implementiert (vgl. **Tabelle 15-10**)

Registriervorgang (mit der Funktion `addPerson`) wird dabei eine neue Person angelegt.

In **Listing 1** ist exemplarisch gezeigt, wie der Modus definiert wird.

Beispiel:

```
<matching-mode>MATCHING_IDENTITIES</matching-mode>
```

Listing 1: XML-Code zum Definieren des Matching-Modes.

15.4.2 mpi-generator

Wird eine Person im E-PIX erstmalig eingetragen, so erhält diese eine MPI-ID. Die Erzeugung einer MPI-ID wird dabei durch einen Generator durchgeführt. Derzeit ist im E-PIX ein Generator (`EAN13Generator`) integriert, welcher eindeutige MPI-IDs erzeugt. Weitere Generatoren können implementiert werden. In **Listing 2** ist die Angabe des Generators dargestellt.

```
<mpi-generator>
      org.emau.icmvc.ttp.epix.gen.impl.EAN13Generator
</mpi-generator>
```

Listing 2: XML-Code zum Definieren des MPI-Generators.

15.4.3 mpi-prefix

Die ersten Ziffern im MPI können mithilfe eines Präfixes festgelegt werden. Jeder MPI enthält damit die angegebene Ziffernfolge[12]. Wird beispielsweise das Präfix *1001* gesetzt, so könnte ein resultierender MPI so aussehen: *1001000000035*. In **Listing 3** ist dargestellt, wie ein Präfix definiert werden kann.

```
<mpi-prefix>1001</mpi-prefix>
```

Listing 3: XML-Code zum Definieren des MPI-Präfixes.

15.4.4 use-notifications

Das Element `use-notifications` dient dazu, bei Änderungen von Datensätzen im E-PIX andere Systeme zu benachrichtigen. Diese Benachrichtigungen werden beispielsweise vom THS-Dispatcher abgerufen. Mit dem Wert `true` wird die Benachrichtigung aktiviert und mit dem Wert `false` deaktiviert. Sind Notifications aktiviert, so werden diese versendet, wenn das Web-Interface verwendet wird. Die SOAP-Schnittstelle stellt für die jeweiligen Methoden eine Variante mit und ohne Versendung von Notifications bereit. Beim Aufruf einer Methode mit Versenden von Notifications, wird in jedem Fall, auch wenn in der Domänenkonfiguration anders definiert, eine Notification versendet. In **Tabelle 15-4** sind alle derzeit unterstützen Notifications aufgelistet.

Tabelle 15-4: Unterstütze Notifications im E-PIX.

[12] Ob das Präfix verwendet wird, hängt davon ab, ob der genutzte MPI-Generator das Präfix berücksichtigt. Der mitgelieferte Generator (`EAN13Generator`) berücksichtigt das Präfix.

Nr.	Name	Beschreibung
1	EPIX.AddIdentifierToPersonNotification	Anfügen eines neuen Identifiers an eine Person.
2	EPIX.AddLocalIdentifierToIdentifierNotification	Anfügen eines neuen lokalen Identifiers an eine Person mit vorhandenen Identifier.
3	EPIX.UpdatePersonNotification	Aktualisierung von Personendaten.
4	EPIX.AddPersonNotification	Person hinzugefügt.
5	EPIX.DeactivatePersonNotification	Person deaktiviert.
6	EPIX.DeletePersonNotification	Person gelöscht.
7	EPIX.SetReferenceIdentityNotification	Identität als Hauptidentität einer Person gesetzt.
8	EPIX.DeactivateIdentityNotification	Identität einer Person deaktiviert.
9	EPIX.DeleteIdentityNotification	Identität einer Person gelöscht.
10	EPIX.AddContactNotification	Kontaktinformation an eine Person angefügt.
11	EPIX.MoveIdentitiesForIdentifierToPersonNotification	Identitäten einer Person mit dem Identifier an eine andere Person übertragen.
12	EPIX.AssignIdentity	Mögliche Dublette zusammen-geführt.

Im **Listing 4** ist beispielhaft die Benachrichtigung deaktiviert.

```
<use-notifications>false</use-notifications>
```

Listing 4: XML-Code zum Aktivieren der Benachrichtigungen über den Dispatcher.

15.4.5 limit-search-to-reduce-memory-consumption

Das Element `limit-search-to-reduce-memory-consumption` dient zur Reduzierung der Belegung des Arbeitsspeichers. Diese Option reduziert den benötigten Arbeitsspeicher, schränkt dafür jedoch die Attribute ein, nach denen eine Person gesucht werden kann. Wenn die Option auf `true` gesetzt wird, dann können die Personen nur anhand der Felder gesucht werden, die auch für das Matching (siehe **Abschnitt 15.4.12.6**) verwendet werden. In **Listing 5** wird exemplarisch das deaktivieren dieser Option dargestellt.

```
<limit-search-to-reduce-memory-consumption>
      false
</limit-search-to-reduce-memory-consumption>
```

Listing 5: XML-Code zum Deaktivieren der Option zur Reduzierung des benötigten Arbeitsspeichers.

15.4.6 persist-mode

Das Element `persist-mode` legt den Modus fest, wie Identitätsdaten gespeichert werden. Dabei kann zwischen `IDENTIFYING` und `PRIVACY_PRESERVING` gewählt werden. Standardmäßig wird (wenn dieses Element nicht angegeben wurde) der Modus `IDENTIFYING` verwendet. Dabei werden alle Daten, die bei der Personenregistrierung übermittelt wurden im E-PIX persistiert. Wird der Modus `PRIVACY_PRESERVING` gewählt, werden alle Daten die nicht einem Ziel-Feld eines Bloomfilters entsprechen, entfernt. Die Daten werden zu keiner Zeit persistiert. Ein Record Linkage kann dann nur auf Basis von Bloomfiltern durchgeführt werden. Weitere Informationen zum Bloomfilter sind unter **Abschnitt 15.4.10** zu finden. In **Listing 6** wird exemplarisch die Festlegung des Persist-Modes dargestellt.

```
<persist-mode>IDENTIFYING</persist-mode>
```

Listing 6: XML-Code zum Wählen des Persist-Modes.

15.4.7 required-fields

Mit dem Element `required-fields` kann festgelegt werden, welche Felder für eine Registrierung verpflichtend übermittelt werden müssen. Eine Auflistung der entsprechenden Felder findet über das Element `name` statt. Eine Auflistung der Feldnamen ist in

Tabelle 15-1 zu finden. In dem nachfolgenden Listing ist exemplarisch eine Konfiguration dargestellt, wodurch zur Registrierung die Felder Vorname, Nachname, Geburtsdatum und Geschlecht übermittelt werden müssen.

```
<required-fields>
      <name>firstName</name>
      <name>lastName</name>
      <name>birthDate</name>
      <name>gender</name>
</required-fields>
```

Listing 7: XML-Code zur Festlegung der Pflichtfelder, die für eine Registrierung übermittelt werden müssen.

15.4.8 value-fields-mapping

Die Felder *value1* – *value10* können für beliebige Werte verwendet werden. Die entsprechenden Felder können mit einem sprechenden Namen versehen werden, welcher in der Weboberfläche (vgl. **Abbildung 15-4**) dargestellt wird. Es handelt sich dabei jedoch nur um ein Label, für etwaige weitere Konfigurationen wird weiterhin der Feldname verwendet. In **Listing 8** wird exemplarisch die Vergabe von Labeln für die Felder *value1* und *value2* dargestellt.

```
<value-fields-mapping>
      <value1>KV-Name</value1>
      <value2>KV-Nummer</value2>
</value-fields-mapping>
```

Listing 8: XML-Code zum Definieren von Labeln für *value*-Felder.

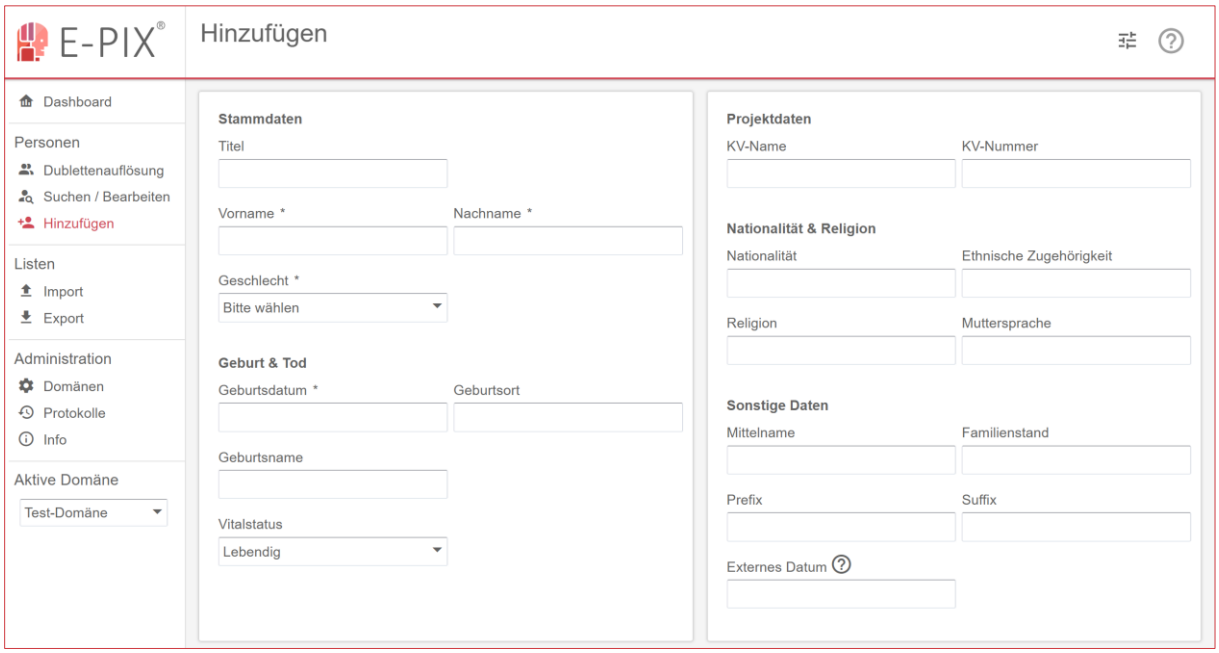


Abbildung 15-4: Weboberfläche zur Registrierung eines Person. Rechts sind die gemappten Felder dargestellt.

15.4.9 Dublettenauflösungsbegründung

Bei einer Dublettenauflösung (vgl. **Abschnitt 7.6**) kann entweder eine Begründung in einem Freitextfeld angegeben werden, oder eine zuvor definierte Begründung ausgewählt werden. Letztere Auswahlmöglichkeiten werden in der Domänenkonfiguration hinterlegt. Im Element `deduplication` kann hierfür eine Liste von Begründungen angelegt werden, welches im Form eines oder mehrerer `reason`-Elemente stattfindet. Jede Begründung erhält einen Namen und eine kurze Beschreibung. In **Listing 9** wird exemplarisch eine Dublettenauflösungsbegründung definiert.

```
<deduplication>
      <reason>
            <name>Tippfehler</name>
            <description>Vertauschte oder fehlende Zeichen</description>
      </reason>
      …
</deduplication>
```

Listing 9: Exemplarisches Beispiel zum Anlegen von Dublettenauflösungsbegründungen. Hier am Beispiel der Begründung „Tippfehler“ mit einer kurzen erklärenden Beschreibung.

15.4.10 privacy

Das Privacy-Element ist ein Container für alle Bloomfilter-Konfigurationen. Der E-PIX unterstützt die Generierung mehrerer Bloomfilter (mittels unterschiedlicher Konfiguration) auf Basis der

Identitätsdaten. Jeder Bloomfilter besteht dabei aus einem `bloomfilter-config`-Element, welches die jeweilige Konfiguration beinhaltet.

15.4.10.1 bloomfilter-config

Die Bloomfilter-Konfiguration enthält alle Einstellungen für einen Bloomfilter. Dabei ist zu beachten, dass 1) das Feld in dem der Bloomfilter gespeichert wird, die Länge des Bloomfilters zulässt (vgl. **Tabelle 15-1**) und 2) der Bloomfilter aus normalisierten bzw. aus aufbereiteten Werten generiert wird (siehe **Abschnitt 15.4.11**). Der Bloomfilter kann wie andere Felder auch zum Matching verwendet werden. Hierzu stehen entsprechende Vergleichsverfahren zur Verfügung. Im **Abschnitt 15.4.12.6.6** sind weitere Informationen dazu enthalten. Zu beachten ist, dass Bloomfilter im E-PIX im base64-Format gespeichert werden.

⚠ **Hinweis:** Der E-PIX unterstützt mehrere Verfahren und zusätzliche Härtungsverfahren, die kombiniert werden können. Achten Sie darauf, dass die Bloomfilter-Konfiguration auf die Bedürfnisse des jeweiligen Projekts angepasst werden muss und so mitunter ein Abwägen zwischen Sicherheit und Qualität stattfinden muss. Ein Bloomfilter stellt immer eine Verallgemeinerung der Eingangsdaten dar und kann zu schlechteren Matching-Ergebnissen führen, sofern der Bloomfilter zum Record Linkage genutzt wird.

In der nachfolgenden Tabelle sind alle Elemente zur Bloomfilter-Konfiguration aufgeführt. Ein exemplarisches Beispiel ist in **Listing 10** aufgeführt.

Tabelle 15-5: Elemente der Bloomfilter-Konfiguration.

Element-Name	Beschreibung	Beispiel
`algorithm`	Angabe des Algorithmus, welcher das Verfahren zur Erzeugung des Bloomfilters implementiert. Eine Auflistung von den unterstützen Algorithmen ist in **Tabelle 15-6** zu finden.	org.emau.icmvc.ttp.deduplication.impl.bloomfilter.RandomHashingStrategy
`field`	Feld der Identität, in dem der Bloomfilter gespeichert werden soll. Dabei zu ist beachten, dass das Feld ggf. überschrieben wird und die Länge des Bloomfilters durch das Feld unterstützt werden muss. Obwohl alle Felder grundsätzlich verwendet werden können, wird die Wahl der Value-Felder 6-8 (**Tabelle 15-1**) empfohlen (je nach Konfiguration).	value8
`length`	Länge des Bloomfilters in Bits.	1000
`ngrams`	Länge der N-Gramme, die für die Erzeugung des Bloomfilters verwendet werden. Klassischerweise	2

	wird hier ein Wert von 2 angegeben, um Bi-Gramme zu erzeugen.	
bits-per-ngram	Anzahl der Bits, die pro N-Gramm im Bloomfilter gesetzt werden. Beim Doube-Hashing wird von Iterationen gesprochen. Beim Random-Hashing handelt es sich um die Anzahl der generierten Zufallspositionen.	25
fold	Der E-PIX unterstützt ein XOR-Folding von Bloomfilern nach Schnell et al.[13]. Der Wert gibt die Anzahl der Faltungen an. Zu beachten ist, dass der Wert+1 ein ganzzahliger Teiler von der Länge des Bloomfilters sein muss ($n+1 \mid Länge$). Wird 0 angegeben, wird der Bloomfilter nicht gefaltet. Pro Faltung halbiert sich die Länge des resultierenden Bloomfilters.	Bei Bloomfilter der Länge 1000 wären möglich: 0, 1, 3, 4, 7, …
alphabet	Das Alphabet, welches beim Random-Hashing berücksichtigt werden soll (nur erforderlich, wenn das Random-Hashing verwendet wird).	ABCDEF12345-.
balanced	Der E-PIX unterstützt das Generieren von Balanced-Bloomfiltern (Schnell et al.[14]). Das Element `balanced` enthält ein Feld `seed`, welches einen Zahlenwert enthält. Dieser stellt den Seed-Wert des Zufallsgenerators dar. Wird dieses Element (`balanced`) nicht angegeben, wird kein Balanced-Bloomfilter erzeugt. Der Balanced-Bloomfilter führt zu einer Verdopplung der resultierenden Bloomfilter-Länge.	462945623209
source-field	Jeder Bloomfilter kann aus einem oder mehreren Feldern zusammengesetzt werden. Dabei wird je Feld (Element: `field` (enthält Feldnamen, siehe **Tabelle 15-1**)) der Wert entsprechend gehashed. Beim Random-Hashing kann pro Feld ein Seed-Wert (Element: `seed` (enthält einen Zahlenwert)) gesetzt werden. Beim Double-Hashing kann ein Salt auf Basis einer statischen Zeichenkette (Element: `salt-value` (enthält eine feste Zeichenkette	

[13] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527984
[14] https://ieeexplore.ieee.org/document/7836669

	(z.B.: a3ghd5o36#sz3)) oder dynamisch auf Basis eines anderen Feldes (Element: salt-field (enthält Feldnamen, siehe **Tabelle 15-1**)) der Identität gesetzt werden.

Der E-PIX unterstützt mehrere Verfahren, um Bloomfilter zu erzeugen. In der nachfolgenden Tabelle sind alle unterstützen Algorithmen angegeben.

Tabelle 15-6: Unterstütze Algorithmen zur Generierung von Bloomfiltern.

Algorithmus	Beschreibung
`org.emau.icmvc.ttp.deduplication.impl.bloomfilter.RandomHashingStrategy`	Random Hashing[15]
`org.emau.icmvc.ttp.deduplication.impl.bloomfilter.DoubleHashingStrategy`	Double Hashing[16]
`org.emau.icmvc.ttp.deduplication.impl.bloomfilter.DoubleHashingStrategyFaster`	Optimierte Variante vom Double Hashing (Nicht Kompatibel mit `DoubleHashingStrategy`)

[15] Schnell R, Borgs C, editors. Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW); 2016 12-15 Dec. 2016.
[16] Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak. 2009;9:41.

```
<privacy>
      <bloomfilter-config>
            <algorithm>org.emau.icmvc.ttp.deduplication.                ↩
                  impl.bloomfilter.RandomHashingStrategy
            </algorithm>
            <field>value8</field>
            <length>1000</length>
            <ngrams>2</ngrams>
            <bits-per-ngram>15</bits-per-ngram>
            <fold>1</fold>
            <alphabet>ABCDEFGHIJKLMNOPQRSTUVWXYZ .-0123456789</alphabet>
            <balanced>
                  <seed>4623829476</seed>
            </balanced>
            <source-field>
                  <name>firstName</name>
                  <seed>456542343</seed>
            </source-field>
            <source-field>
                  <name>lastName</name>
                  <seed>374027465</seed>
            </source-field>
      </bloomfilter-config>
      <bloomfilter-config>
            <algorithm>org.emau.icmvc.ttp.deduplication.                ↩
                  impl.bloomfilter.DoubleHashingStrategy
            </algorithm>
            <field>value6</field>
            <length>500</length>
            <ngrams>2</ngrams>
            <bits-per-ngram>15</bits-per-ngram>
            <source-field>
                  <name>firstName</name>
                  <salt-field>birthDate</salt-field>
            </source-field>
            <source-field>
                  <name>gender</name>
                  <salt-value>Q2fh-Fk2#CjP+s5#</salt-value>
            </source-field>
      </bloomfilter-config>
</privacy>
```

Listing 10: Verkürzte exemplarische Konfiguration von zwei Bloomfiltern.

15.4.11 preprocessing-config

Mithilfe des pre-prossessings können Felder aufbereitet werden. Dies ermöglicht beispielsweise, dass für das Record Linkage z.B. die Vornamen ohne Berücksichtigung der Groß- und Kleinschreibung miteinander verglichen werden. Ein pre-processing muss maximal für die Felder durchgeführt werden, die beim Record Linkage verwendet werden. Die Felder werden in jedem Fall im unbearbeiteten Zustand, demnach so wie diese übermittelt wurden, im E-PIX abgelegt.

Im Element `preprocessing-config` werden alle `preprocessing-fields` aufgelistet. In **Listing 11** ist ein einfaches Beispiel für die Konfiguration der Aufbereitung des Feldes für den Vornamen.

```
<preprocessing-config>
      <preprocessing-field>
            <field-name>firstName</field-name>
            <simple-transformation-type                                    ↩
              xsi:type="ma:SimpleTransformation">
                <input-pattern> </input-pattern>
                <output-pattern></output-pattern>
            </simple-transformation-type>
            <complex-transformation-type                                   ↩
              xsi:type="ma:ComplexTransformation">
                <qualified-class-name>org.emau.icmvc.ttp.                  ↩
                  deduplication.preprocessing.impl.                        ↩
                  ToUpperCaseTransformation                                ↩
                </qualified-class-name>
            </complex-transformation-type>
      </preprocessing-field>
<preprocessing-config>
```

Listing 11: Exemplarischer XML-Code mit allen Elementen für ein pre-processing eines Feldes.

In **Abschnitt 15.4.11.2** wird das Element `field-name`, in **Abschnitt 15.4.11.3** wird das Element `simple-transformation-type` und in **Abschnitt 15.4.11.4** das Element `complex-transformation-type` erläutert.

15.4.11.1 preprocessing-field

Im Element `preprocessing-field` ist zum einen das betroffene Feld angegeben und alle Transformationen, die für die Aufbereitung eines Feldes verwendet werden sollen. Dabei wird zwischen einfachen und komplexen Transformationen unterschieden, die sich jeweils in ihrer Konfiguration unterscheiden. Eine einfache Transformation stellt ein einfaches Ersetzen dar. Hierbei wird eine bestimmte Zeichenkette in einem Feld gesucht und durch eine andere Zeichenkette ersetzt. Eine komplexe Transformation bezieht sich auf den Inhalt eines gesamten Feldes. Die durchgeführte Operation hängt dabei von der verwendeten Transformation ab.

⚠ **Hinweis:** Die Reihenfolge der Transformationen ist nicht sichergestellt und kann von der Reihenfolge der Definition in der XML-Datei abweichen. `complex-transformation-type` werden stets nach `simple-transformation-type` verarbeitet[17].

15.4.11.2 field-name

Das Element `field-name` gibt das Feld an, welches aufbereitet werden soll. Die Bezeichnungen für die jeweiligen Felder sind in

Tabelle 15-1 angegeben.

[17] Die Festlegung der Reihenfolge wird demnächst implementiert.

15.4.11.3 simple-transformation-type

Mithilfe des Elements `simple-transformation-type` kann eine definierte Zeichenkette durch eine andere ersetzt werden. Hierzu wird mittels des Elements `input-pattern` die Zeichenkette definiert, die ersetzt werden soll. Mit dem Element `output-pattern` kann die Zeichenkette angegeben werden, die eingefügt wird. Diese kann auch leer sein, dann wird die gefundene Zeichenkette nur entfernt. In **Listing 12** sind zwei `simple-transformation-type` dargestellt. Die erste Transformation dient zum Entfernen von allen Leerzeichen aus einem Feld, die Zweite ersetzt das Zeichen *é* durch *e*.

```
...
<simple-transformation-type xsi:type="ma:SimpleTransformation">
      <input-pattern> </input-pattern>
      <output-pattern></output-pattern>
</simple-transformation-type>
<simple-transformation-type xsi:type="ma:SimpleTransformation">
      <input-pattern>é</input-pattern>
      <output-pattern>e</output-pattern>
</simple-transformation-type>
...
```

Listing 12: XML-Code zur Definition zweier einfacher Transformationen.

15.4.11.4 complex-transformation-type

Mithilfe des Elements `complex-transformation-type` kann eine Transformation auf ein gesamtes Feld angewendet werden. Dies bedeutet nicht, dass alle Zeichen betroffen sind. Welche Transformation angewendet werden soll, wird mithilfe des Elements `qualified-class-name` angegeben. Die derzeit implementierten Transformationen sind in **Tabelle 15-7** genannt und beschrieben. Dabei ist zu beachten, dass bei der Angabe der Transformation immer noch `org.emau.icmvc.ttp.deduplication.preprocessing.impl.` vorangestellt werden muss.

Tabelle 15-7: Unterstützte Transformationen für `complex-transformation-type`.

Transformation	Beschreibung	Beispiel
`ToUpperCaseTransformation`	Alle Kleinbuchstaben werden durch Großbuchstaben ersetzt.	Anna ➔ ANNA
`CharsMutationTransformation`	Ersetzt Umlaute.	München ➔ Muenchen
`TrimTransformation`	Entfernt führende und nachfolgende Leerzeichen.	␣␣An␣na␣ ➔ An␣na
`CharNormalization Transformation` ↩	Normalisiert alle Zeichen nach ASCII	â ➔ a é ➔ e

In **Listing 13** wird exemplarisch gezeigt, wie führende und nachfolgende Leerzeichen für das Record Linkage mittels Transformator entfernt werden.

```
...
<complex-transformation-type xsi:type="ma:ComplexTransformation">
      <qualified-class-name>
            org.emau.icmvc.ttp.deduplication.preprocessing.impl.        ↩
            TrimTransformation
      </qualified-class-name>
</complex-transformation-type>
...
```

Listing 13: XML-Code zur Definition eines Transformators zum Entfernen führender und nachfolgender Leerzeichen.

15.4.12 matching

Das Record Linkage wird mithilfe des Elements `matching` konfiguriert. Im E-PIX wird das Verfahren von Fellegi-Sunter zur Bestimmung von Wahrscheinlichkeiten verwendet. Hierzu werden die Felder konfiguriert, welche für das Blocking und das Matching verwendet werden sollen. Mithilfe von zwei Schwellwerten (`threshold-possible-match` und `threshold-automatic-match`) kann zwischen 4 Match-Typen unterschieden werden. Die Schwellwerte können dem Verfahren entsprechend angepasst werden. Werden die Elemente nicht angegeben, werden Standardwerte gesetzt. In **Tabelle 15-8** sind die empfohlenen und Standard-Schwellwerte dargestellt.

Tabelle 15-8: Empfohlene und Standard-Schwellwerte für *Automatic Match* und *Possible Match*.

Schwellwert	Wert	Standardwert
`threshold-automatic-match`	14,5	20
`threshold-possible-match`	2,99	4

Die Match-Typen wurden In **Abschnitt 7.2** erläutert. In **Tabelle 7-1** sind alle Match-Typen aufgeführt und entsprechend erläutert.

15.4.12.1 threshold-possible-match

Mit dem Element `threshold-possible-match` kann der Schwellwert für *Possible Matches* (vgl. **Tabelle 15-8**) definiert werden. Überschreitet die ermittelte Wahrscheinlichkeit den angegeben Wert (und unterschreitet den Schwellwert `threshold-automatic-match`), so wird der Match-Typ *Possible Match* als Ergebnis des Record Linkages zurückgegeben. In **Listing 14** ist die Definition des Schwellwert dargestellt.

```
<threshold-possible-match>2.99</threshold-possible-match>
```

Listing 14: XML-Code zur Definition des Schwellwerts für *Possible Matches*.

15.4.12.2 threshold-automatic-match

Mit dem Element `threshold-automatic-match` kann der Schwellwert für *Automatic Matches* (vgl. **Listing 15**) definiert werden. Unterscheiden sich die abgeglichenen Datensätze voneinander, die ermittelte Wahrscheinlichkeit überschreitet jedoch den angegebenen Wert, so wird der Match-Typ

Automatic Match als Ergebnis des Record Linkages zurückgegeben. In **Listing 15** ist die Definition des Schwellwert dargestellt.

```
<threshold-automatic-match>14.5</threshold-automatic-match>
```

Listing 15: XML-Code zur Definition des Schwellwerts für *Automatic Matches*.

ⓘ Wie können automatische Zusammenführungen (*Automatische Matches*) deaktiviert werden?

Eine automatische Zusammenführung kann auf *Perfect Matches* beschränkt werden. Fälle mit sehr höher Übersteinstimmung, die trotz kleiner Unterschiede zusammengeführt werden würden (*Automatischer Match*), können somit manuell geprüft werden. Hierzu wird der Schwellwert für `threshold-automatic-match` auf 1001 gesetzt. Damit liegt diese über dem internen Wert für *Perfect Matches* (von 1000) und wird so niemals „vor“ einem *Perfect Match* erreicht.

15.4.12.3 use-cemfim

CEMFIM steht für *Check Equal Match for Identifier Match* und dient dazu das Matchingergebnis zu beeinflussen. Dabei kann definiert werden, wie sich der E-PIX verhalten soll, wenn ein übermittelter Identifier mit dem einer Identität übereinstimmt, jedoch mindestens ein Match mit einer Identität einer anderen Person vorhanden ist. Das Element kann die Werte `true` oder `false` annehmen. Das Verhalten des E-PIX kann aus **Tabelle 15-9** entnommen werden.

Tabelle 15-9: Verhalten des E-PIX, je nachdem wie das Element `use-cemfim` definiert wurde.

Nummer	CEMFIM	Mehr als 1 Match vorhanden (mit anderer Person)	Verhalten
1	`true`	Ja	Fehler: Ein Identifier darf nur einer Person pro Domäne zugeordnet sein.
2	`false`	Ja	Die Identität wird gespeichert und als Possible Match hinterlegt.
3	`true`	Nein	Die Identität wird gespeichert und als Possible Match hinterlegt.
4	`false`	Nein	Die Identität wird gespeichert und als Possible Match hinterlegt.

In **Listing 16** ist exemplarisch die Definition dargestellt.

```
<use-cemfim>true</use-cemfim>
```

Listing 16: XML-Code zur Definition des *use-cemfim*-Wertes.

15.4.12.4 parallel-matching-after

Der E-PIX unterstützt Multithreading, wodurch die Performance gesteigert wird. Bei einer niedrigen Anzahl von registrierten Identitäten ist es performanter einen sequenziellen Abgleich durchzuführen. Deshalb kann mit dem Element `parallel-matching-after` definiert werden, ab wieviel

registrierten Identitäten ein paralleler Abgleich, also verteilt auf mehrere Threads, stattfinden soll. Der Wert ist abhängig von der Rechenleistung des Systems. Bei einem erwarteten Datenbestand von mehreren Tausend registrierten Identitäten sollte der Wert nicht zu hoch gewählt werden. Wird der Wert nicht definiert, so wird standardmäßig 1000 gesetzt. In **Listing 17** ist exemplarisch die Definition dargestellt.

```
<parallel-matching-after>1000</parallel-matching-after>
```

Listing 17: XML-Code zur Definition der Anzahl registrierter Personen, ab denen der E-PIX Multithreading verwendet.

15.4.12.5 number-of-threads-for-matching

Die Anzahl der verwendeten Threads kann definiert werden. Dabei wird diese in Abhängigkeit des verwendeten Systems eingestellt. Wenn das Element nicht definiert wird, liegt der Wert standardmäßig bei 4 Threads. Je nachdem, wie viele Threads der E-PIX verwenden soll, kann der Wert erhöht oder verringert werden. Eine höhere Anzahl von Threads bedeutet, dass im Optimalfall ein Abgleich von Personen schneller durchgeführt werden kann, da die Vergleiche auf mehrere Threads aufgeteilt werden. Insbesondere bei großen Datenbeständen kann eine Verteilung auf mehrere Threads deutlich performanter sein. In **Listing 18** ist die exemplarische Definition der Anzahl der verwendeten Threads dargestellt.

```
<number-of-threads-for-matching>4</number-of-threads-for-matching>
```

Listing 18: XML-Code zur Definition der Anzahl der verwendeten Threads.

15.4.12.6 field

Mit dem Element `field` werden alle Felder definiert, die im Rahmen des Blockings oder/und Matchings verwendet werden. Jedes Feld wird hierfür separat konfiguriert. In **Listing 19** ist exemplarisch angegeben, wie eine Konfiguration eines Feldes aussehen kann. Im Folgenden werden die einzelnen Elemente erläutert.

```
<field>
      <name>gender</name>
      <matching-threshold>0.75</matching-threshold>
      <weight>3</weight>
      <algorithm>
            org.emau.icmvc.ttp.deduplication.impl.LevenshteinAlgorithm
      </algorithm>
</field>
```

Listing 19: XML-Code zur exemplarischen Konfiguration eines Felders, welches zum Matching verwendet wird.

15.4.12.6.1 name

Das Element `name` gibt an, welches Feld für das Blocking oder/und Matching verwendet werden soll. Die Bezeichnungen für die jeweiligen Felder sind in

Tabelle 15-1 angegeben. In **Listing 20** ist exemplarisch der Wert „gender“ angegeben, wenn das Geschlecht z.B. für das Blocking verwendet werden soll.

```
<name>gender</name>
```

Listing 20: XML-Code zur Definition des Feldes für das Record Linkage.

15.4.12.6.2 blocking-threshold

Beim Blocking wird ein erster Abgleich durchgeführt, um eine erste Selektierung durchzuführen. Die Schwellwerte sollten hierfür niedriger angesetzt werden, damit potentielle Duplikate nicht aufgrund eines Abgleichs mit reduzierter Anzahl von abgeglichenen Feldern aussortiert werden. Wird keine entsprechende Schwelle gesetzt, wird standardmäßig der Wert 0.0 gesetzt. Dieser Schwellwert besitzt einen Wertebereich von 0.0 bis 1.0, wobei 0.0 als 0% und 1.0 als 100% zu verstehen ist. In **Listing 21** wird exemplarisch ein Schwellwert definiert.

```
<blocking-threshold>0.8</blocking-threshold>
```

Listing 21: XML-Code zur Definition eines Schwellwertes für das Blocking von einem Feld.

15.4.12.6.3 blocking-mode

Das Blocking unterstützt zwei Datentypen für einen Abgleich zweier Felder. Zum einen `TEXT`, für beliebige Zeichenketten und `NUMBERS` für Zahlen. Letzteres stellt für Zahlen eine Optimierung dar und ist performanter. Dies kann beispielsweise beim Feld Geburtsdatum verwendet werden. Wenn das Element `blocking-mode` nicht angegeben wurde, wird standardmäßig `TEXT` verwendet. In **Listing 22** ist die Definition von `blocking-mode` exemplarisch für Zahlenvergleiche dargestellt.

```
<blocking-mode>NUMBERS</blocking-mode>
```

Listing 22: XML-Code zur Definition der Blocking-Vergleichsmethode.

15.4.12.6.4 matching-threshold

Ist beim Matching der ermittelte Wert der Übereinstimmung gleich oder höher dem im Element `matching-threshold` definierten Wert, dann liegt ein Match für das entsprechende Feld vor. Anders als beim Blocking sollte der Schwellwert höher angesetzt werden, weil beim Matching nur tatsächliche Duplikate ermittelt werden sollen. Trotzdem sollte der Schwellwert genug Raum für etwaige Fehler (z.B. Tippfehler, Zahlendreher) lassen, damit beim Abgleich diese dennoch als Duplikate erkannt werden können. Der Schwellwert hängt von dem entsprechenden Feld ab und muss dementsprechend an das Feld angepasst werden. Der Schwellwert besitzt einen Wertebereich von 0.0 bis 1.0, wobei 0.0 als 0% und 1.0 als 100% zu verstehen ist. In **Listing 23** ist exemplarisch eine Schwelle definiert.

```
<matching-threshold>0.8</matching-threshold>
```

Listing 23: XML-Code zur Definition eines Schwellwertes für das Matching von einem Feld.

15.4.12.6.5 weight

Mit dem Element `weight` kann eine Wichtung definiert werden. Damit kann bestimmt werden, wie sehr das Ergebnis eines Vergleichs in das Gesamtergebnis einfließt. Je höher der Wert ist, desto höher gewichtet wird das Feld. Wenn kein Wert angegeben wurde, wird der Wert 1 standardmäßig verwendet. In **Listing 24** ist exemplarisch eine Wichtung angegeben.

```
<weight>3</weight>
```

Listing 24: XML-Code zur Wichtung eines Feldes.

15.4.12.6.6 algorithm

Der Abgleich der Felder kann mittels unterschiedlicher Verfahren durchgeführt werden. Hierfür wird im Element `algorithm` der Algorithmus eingetragen, welcher für das Matching verwendet werden soll. In **Tabelle 15-10** sind alle derzeit unterstützten Verfahren aufgelistet und erläutert. Bei der Angabe des Algorithmus muss immer ein `org.emau.icmvc.ttp.deduplication.impl.` vorangestellt werden.

Tabelle 15-10: Unterstütze Algorithmen für das Matching.

Algorithmus	Beschreibung
`ColognePhoneticAlgorithm`	Vergleicht zwei Werte nach ihrem Sprachklang. Die Nachnamen Maier, Meyer und Meier würden beispielsweise als gleich gewertet werden.
`DeterministicAlgorithm`	Vergleicht zwei Werte auf exakte Gleichheit. Bei exakter Gleichheit zweier Werte ist das Ergebnis 1, bei einer Abweichung 0.
`LevenshteinAlgorithm`	Vergleicht zwei Werte anhand ihrer Levenshtein-Distanz. Dabei werden durch Einfügen oder Löschen von Zeichen zwei Zeichenketten aneinander angeglichen. Je weniger Operationen nötig sind, desto Ähnlicher sind sich zwei Werte. Dies stellt die empfohlene Methode für das Matching dar und wird standardmäßig verwendet.
`SorensenDiceCoefficientCoded`	Vergleicht zwei (base64-kodierte) Bloomfilter auf Ähnlichkeit. Dabei wird der Sørensen-Dice Koeffizient verwendet. Dieser Algorithmus wird verwendet, wenn der Bloomfilter direkt im E-PIX erzeugt wurde.
`JaccardSimilarityAlgorithmCoded`	Vergleicht zwei (base64-kodierte) Bloomfilter auf Ähnlichkeit. Dabei wird der Jaccard-Koeffizient verwendet. Dieser Algorithmus wird verwendet, wenn der Bloomfilter direkt im E-PIX erzeugt wurde.
`SorensenDiceCoefficient`	Vergleicht zwei (0 und 1 basierte String-) Bloomfilter auf Ähnlichkeit. Dabei wird der Sørensen-Dice Koeffizient verwendet. Dieser Algorithmus wird verwendet, wenn der Bloomfilter

	nicht im Base64-Format dem E-PIX übermittelt wird, sondern in einem 0 und 1 basierten String vorliegt.
`JaccardSimilarityAlgorithm`	Vergleicht zwei (0 und 1 basierte String-) Bloomfilter auf Ähnlichkeit. Dabei wird der Jaccard-Koeffizient verwendet. Dieser Algorithmus wird verwendet, wenn der Bloomfilter nicht im Base64-Format dem E-PIX übermittelt wird, sondern in einem 0 und 1 basierten String vorliegt.

In **Listing 25** wird exemplarisch die Definition eines Algorithmus zum Abgleich von einem Feld definiert.

```
<algorithm>
      org.emau.icmvc.ttp.deduplication.impl.LevenshteinAlgorithm
</algorithm>
```

Listing 25: XML-Code zur Definition des Algorithmus für das Matching.

15.4.12.7 multiple-values

Der E-PIX unterstützt sogenannte Multiple-Value-Fields. Hierbei werden Teil-Zeichenketten innerhalb eines Feldes in unterschiedlichen Reihenfolgen abgeglichen. Sind beispielsweise mehrere Vornamen innerhalb des Feldes *Vorname* angegeben, so werden bei einem Vergleich alle Permutationen der Reihenfolgen abgeglichen. Es wäre somit beispielsweise irrelevant, ob eine Person den Vornamen mit „Klaus Dieter“ oder „Dieter Klaus“ angibt[18]. Hierzu kann ein Separator definiert werden, anhand dessen die Teil-Zeichenketten ermittelt werden. In **Listing 26** ist exemplarisch ein `multi-value` dargestellt. Die enthaltenen Elemente werden im Folgenden erläutert.

```
<multiple-values>
      <separator> </separator>
      <penalty-not-a-perfect-match>0.1</penalty-not-a-perfect-match>
      <penalty-one-short>0.1</penalty-one-short>
      <penalty-both-short>0.2</penalty-both-short>
</multiple-values>
```

Listing 26: XML-Code zur Definition eines `multi-value`-Feldes.

15.4.12.7.1 separator

Mit dem Element `separator` kann ein Zeichen definiert werden, anhand dessen ein Wert in mehrere Zeichenketten aufgespalten wird. Beim Feld Vorname könnte dies beispielweise ein Leerzeichen sein, sodass sich z.B. aus „Klaus Dieter“ die Teil-Zeichenketten „Klaus“ und „Dieter“ ergeben. Ein Abgleich findet dann unabhängig der Reihenfolge der Teil-Zeichenketten statt. Zu beachten ist, dass nur ein Zeichen als Separator dienen kann. In **Listing 27** ist die Definition eines Leerzeichens als Separator dargestellt.

[18] Dabei ist entscheidend, welcher Separator verwendet wird.

```
<separator> </separator>
```

Listing 27: XML-Code zur exemplarischen Definition eines Leerzeichens als Separator eines multi-value-Feldes.

15.4.12.7.2 penalty-not-a-perfect-match

Mit dem Element `penalty-not-a-perfect-match` kann ein Wert definiert werden, der dem Wahrscheinlichkeitswert einer Übereinstimmung abgezogen wird, wenn bei einem Multiple Value Feld zwar alle Teilzeichenketten matchen, aber nicht exakt gleich sind. Beispiel: Klaus Dieter und Klaas Dieter. Klaas und Klaus sind ähnlich genug und matchen daher, sie unterscheiden sich jedoch geringfügig. In **Listing 29** ist exemplarisch gezeigt, wie ein Wert hierfür definiert wird.

15.4.12.7.3 penalty-one-short

Mit dem Element `penalty-one-short` kann ein Wert definiert werden, der dem Wahrscheinlichkeitswert einer Übereinstimmung abgezogen wird, wenn bei einem Multiple Value Feld nicht alle Teilzeichenketten matchen. Beispiel: Klaus Dieter und Klaus. Klaus matcht, Dieter fehlt jedoch in einem Feld.

```
<penalty-one-short>0.1</penalty-one-short>
```

Listing 28: XML-Code zur exemplarischen Definition des *penalty-one-short*-Wertes.

15.4.12.7.4 penalty-both-short

Mit dem Element `penalty-both-short` kann ein Wert definiert werden, der dem Wahrscheinlichkeitswert einer Übereinstimmung abgezogen wird, wenn bei beiden Multiple Value Feldern nicht alle Teilzeichenketten matchen. Beispiel: Klaus Dieter und Dieter Erhardt. In **Listing 29** ist exemplarisch gezeigt, wie ein Wert hierfür definiert wird.

```
<penalty-both-short>0.2</penalty-both-short>
```

Listing 29: XML-Code zur exemplarischen Definition des *penalty-both-short*-Wertes.

16 Optimierungen

16.1 Optimierungen bei Multi-Millionen Beständen

Bei Datenbeständen mit mehreren Millionen zu verwaltenden Personen, können in Abhängigkeit der Leistungsfähigkeit der verwendeten Hardware, höhere Laufzeiten entstehen. Dies kann es erforderlich machen, weitere Anpassungen vorzunehmen. Diese sollten aber ausdrücklich erst dann vorgenommen werden, wenn entsprechende Datenbestände erreicht oder erwartet werden. Dies umfasst beispielsweise das Hochsetzen von Timeouts, was nur bedingt durch den Datenbestand sinnvoll ist, aber nicht grundsätzlich.

1. Wert für Timeout in der Datenbank erhöhen

Bei großen Datenmengen können die standardmäßigen Zeiten bis zum Auslösen von Timeouts zu niedrig sein. Treten diese auf, so können diese in der Datenbank erhöhrt werden. Hier wird muss die (Datenbank-)Servervariable `innodb_lock_wait_timeout` erhöht werden. Standardmäßig liegt diese bei 50 Sekunden.

2. Werte für Timeout des WildFly Applikationsservers erhöhen

Wenn der Start eines Deployments zu lange dauert (standardmäßig mehr als 5 Min.), dann wird ein Timeout ausgelöst. Beim E-PIX kann das passieren, wenn der Datenbestand groß ist und nicht schnell genug alle Daten aus der Datenbank in den Cache geladen werden können. Dieser Abschnitt kann hierzu in die Konfiguration des Applikationservers WildFly eingefügt und der Wert angepasst werden:

```
<system-properties>
      <property name="jboss.as.management.blocking.timeout"          ↩
            value="DAUER IN SEKUNDEN" />
</system-properties>
```

Gleiches gilt für die Demployment-Dauer (standardmäßig 60 Sekunden). Folgende bereits vorhandene Konfiguration muss dafür angepasst werden:

```
<subsystem xmlns="urn:jboss:domain:deployment-scanner:2.0">
      <deployment-scanner deployment-timeout="DAUER IN SEKUNDEN"      ↩
      ... />
</subsystem>
```

Der WildFly definiert selbst eine maximale Dauer für Datenbankabfragen. Diese muss auch angepasst werden (z.B. wenn die Anpassung unter 1. überstiegen wird). Standardmäßig wird dieser Timeout nach 300 Sekunden ausgelöst. Die bereits enthaltene Konfiguration muss entsprechend angepasst werden:

```
<subsystem xmlns="urn:jboss:domain:transactions:3.0">
      ...
      <coordinator-environment default-timeout="DAUER IN SEKUNDEN" />
</subsystem>
```

16.2 Optimierungen bei Betrieb ohne Docker

Wird entgegen der hier beschrieben Vorgehensweise selbst ein Applikationsserver und Datenbankserver aufgesetzt, so kann eine Performance-Steigerung des E-PIX® durch diverse Optimierungen erzielt werden. In den von der Treuhandstelle Greifswald ausgelieferten Docker-Containern (WildFly und MySQL) sind diese bereits eingerichtet. Diese Optimierungen sind relevant, wenn größere Datenbestände mit mehreren 10-Tausend Personen erwartet werden.

16.2.1 Speicher für MySQL erhöhen

Standardmäßig ist im MySQL-Server eine `innodb_buffer_pool_size` von 128 MB eingestellt. Es wird empfohlen diese auf 2 GB zu erhöhen. Dies geschieht entweder direkt in der Datenbank oder bei der Verwendung eines Docker-Containers als entsprechendes Kommando. Bei der Konfiguration dieses Wertes ist die offizielle MySQL-Dokumentation

(https://dev.mysql.com/doc/refman/5.7/en/innodb-buffer-pool-resize.html) zu beachten. Die Anpassung dieses Wertes erfolgt unter Beachtung des verfügbaren Arbeitsspeichers.

16.2.2 Batch-Writing

Für jede Datenbankoperation (Insert, Update, Delete) wird standardmäßig separat auf die Datenbank zugegriffen. Zur Steigerung der Performance können die Anfragen jedoch zusammengefasst werden. Dies kann erreicht werden, indem in der `standalone.xml` des WildFly-Servers der Parameter `rewriteBatchedStatements=true` an die `jdbc-connection-url` angefügt wird.

16.2.3 Lange Zeiten zum Hochfahren des Applikationsservers

Wurden viele Millionen Pseudonyme angelegt und ein Neustart des Systems ist erforderlich, so kann das Hochfahren des Applikationsservers WildFly mehr Zeit in Anspruch nehmen, als der konfigurierte Timeout zulässt. Der Timeout wird standardmäßig nach 5 Minuten ausgelöst, sofern der WildFly bis dahin nicht hochgefahren ist. Es ist dann erforderlich, die Konfiguration des WildFly abzupassen. Hierzu wird in der `standalone.xml` des WildFly-Servers die Komponente `deployment-scanner` um das Attribut `deployment-timeout` ergänzt. Der Wert des Attributes gibt die Zeit in Sekunden an, ab wann ein Timeout ausgelöst wird. Im folgenden Beispiel wird der Timeout auf 15 Minuten (900 Sekunden) hochgesetzt.

```
<subsystem xmlns="urn:jboss:domain:deployment-scanner:2.0">
    <deployment-scanner […] scan-interval="5000"
    deployment-timeout="900" […]/>
</subsystem>
```

17 Publikationen und Vorträge

- Hampf C, Geidel L, Zerbe N, Bialke M, Stahl D, Blumentritt A, Bahls T, Hufnagl P, Hoffmann W, et al.
  Assessment of scalability and performance of the record linkage tool E-PIX® in managing multi-million patients in research projects at a large university hospital in Germany (Originalartikel)
  Journal of Translational Medicine. 2020. DOI:10.1186/s12967-020-02257-4
  https://dx.doi.org/10.1186/s12967-020-02257-4

- Bialke M*, Penndorf P, Wegner T, Bahls T, Havemann C, Piegsa J, et al.
  A workflow-driven approach to integrate generic software modules in a Trusted Third Party (Originalartikel)
  Journal of Translational Medicine. 2015; 13(176).
  http://www.translational-medicine.com/content/13/1/176

- Bialke M*, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, et al.
  MOSAIC. A modular approach to data management in epidemiological studies. (Originalartikel)
  METHODS OF INFORMATION IN MEDICINE. 2015; 54(4):364-371.
  http://dx.doi.org/10.3414/ME14-01-0133

- Bialke M, Langner D, Bahls T, Geidel L, Piegsa J, Havemann C, Hoffmann W.
  “Who am I? And if so, how many?” – The E-PIX as innovative system to manage person identities. (Poster)
  2nd Research Data Management Workshop; 2014 Nov 27; Köln.

18 Weiterführende Informationen

Überblicksseite E-PIX und Download

https://www.ths-greifswald.de/epix

Produktbroschüre E-PIX

https://www.ths-greifswald.de/epix/produktbrief

E-PIX Service Spezifikation

https://www.ths-greifswald.de/epix/doc

E-PIX Demo

https://www.ths-greifswald.de/epix/demo

Offizielles E-PIX Docker-Image

https://www.ths-greifswald.de/forscher/e-pix/#download

Git-Repository (Stand MOSIAC-Projekt, aktuelle Version siehe Offizielles Docker-Image)

https://github.com/mosaic-hgw/E-PIX

Docker Installation

https://docs.docker.com/install/

Docker-Compose Installation

https://docs.docker.com/compose/install/

Docker Cheat Sheet

https://www.docker.com/sites/default/files/Docker_CheatSheet_08.09.2016_0.pdf

Docker und Docker-Compose Cheat Sheet

https://dev-eole.ac-dijon.fr/doc/cheatsheets/docker.html