

Anwenderhandbuch



Enterprise Identifier Cross-Referencing

Version für E-PIX 2023.1.* vom 05.05.2023

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

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
2023.1	05.05.2023	Aktualisierung auf E-PIX Version 2023.1 Ergänzung zu Domänen-spezifischen Rollen	Christopher Hampf

Inhalt

Anwenderhandbuch	1
Versionierung	2
Inhalt	3
Abbildungsverzeichnis	4
Tabellenverzeichnis	5
1 Hintergrund	6
2 Der Enterprise Identifier Cross-Referencing (E-PIX)	7
3 Begriffsbestimmungen	7
4 Das Konzept der Nebenidentitäten	9
5 Funktionalitäten	10
5.1 Was leistet der Dienst.....	10
5.2 Was leistet der Dienst nicht.....	10
6 Installation per Docker	11
6.1 Systemanforderungen	11
6.2 Download und Starten des Dienstes.....	11
7 Die grafische Benutzeroberfläche des E-PIX	13
7.1 Anlegen von Domänen, Quellen und Identifier-Domänen	13
7.2 Registrierung einer Person	22
7.3 Suche einer Person anhand demografischer Informationen.....	24
7.4 Einsehen von Details zu einer Person.....	25
7.5 Bearbeiten und Löschen von Personendaten.....	26
7.6 Dublettenauflösung	27
7.7 Daten exportieren.....	28
7.8 Daten importieren	29
7.9 Einsehen von Protokollen	31
7.10 Einsehen von Statistiken mittels des Dashboards	32
8 Logging	33
9 Versand von Notifications	33
10 FHIR-Unterstützung für E-PIX per TTP-FHIR Gateway	33
11 Authentifizierung und Autorisierung	35
11.1 Global.....	35
11.2 Domänen-spezifische Rollen mit OpenID-Connect (Keycloak).....	37
12 Empfehlungen zur Absicherung des Anwendungsservers	37
13 Nutzung der SOAP-Schnittstelle	37

13.1	Registrierung von Personen.....	38
13.2	Personen per MPI suchen.....	39
13.3	Alle Personendaten zu einer Domain	40
14	Konfiguration von E-PIX Domänen	40
14.1	Hintergrund.....	41
14.2	XML-basierte Konfiguration.....	42
14.3	Die Standard-Konfiguration	45
14.4	Struktur und Inhalt der Konfiguration	46
15	Optimierungen	64
15.1	Optimierungen bei Multi-Millionen Beständen.....	64
15.2	Optimierungen bei Betrieb ohne Docker.....	65
16	Publikationen und Vorträge	66
17	Weiterführende Informationen	67

Abbildungsverzeichnis

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.....	6
Abbildung 6-1:	Architektur des E-PIX® mit Docker.....	12
Abbildung 7-1:	Oberfläche zum Anlegen von Domänen, Quellen und Identifier-Domänen.....	14
Abbildung 7-2:	Oberfläche zum Eintragen von Personendaten.	23
Abbildung 7-3:	Oberfläche zum Suchen von Personen anhand von demographischen Daten.....	25
Abbildung 7-4:	Detailseite zur Einsicht von den Stammdaten einer Person.....	25
Abbildung 7-5:	Oberfläche zum Bearbeiten der Stammdaten einer Person.....	27
Abbildung 7-6:	Gegenüberstellung von Personendaten zum Auflösen einer Dublette.	28
Abbildung 7-7:	Oberfläche zum Exportieren von Personendaten.....	29
Abbildung 7-8:	Oberfläche zum Importieren von Personendaten.	29
Abbildung 7-9:	Oberfläche mit Vorschau der ersten eingelesenen Zeilen.....	30
Abbildung 7-10:	Oberfläche zum Einsehen des Protokolls.....	32
Abbildung 7-11:	Dashboard zum Einsehen der Statistiken.	32
Abbildung 13-1:	Exemplarische Anfrage zur Registrierung einer Person.....	38

Abbildung 13-2: Gekürzte Antwort auf die Anfrage zur Registrierung einer Person.	39
Abbildung 13-3: Exemplarische Anfrage zum Suchen einer Person mittels dazugehörigen MPI.	39
Abbildung 13-4: Antwort auf die Anfrage zum Suchen einer Person mittels des MPIs.	40
Abbildung 13-5: Anfrage um alle Personendaten einer Domain abzurufen.	40
Abbildung 14-1: Vereinfachter Ablauf des Matching-Prozesses.	42
Abbildung 14-2: Das Anzeigen und Editieren der aktuellen Konfiguration einer E-PIX-Domäne ist direkt über das Web-Frontend möglich.	43
Abbildung 14-3: Alle Elemente, die bei der Konfiguration der Domäne verwendet werden können.	43
Abbildung 14-4: Weboberfläche zur Registrierung einer Person. Rechts sind die gemappten Felder dargestellt.	50

Tabellenverzeichnis

Tabelle 7-1: Mögliche Match-Typen	24
Tabelle 11-1: Nutzer der Gruppe Admin und User haben unterschiedliche Zugriffsrechte in der Web-Oberfläche.	36
Tabelle 14-1: Alle im E-PIX definierten Felder.	44
Tabelle 14-2: Verwendete Felder mit Schwellwerten und Wichtung in der Standard-Domänenkonfiguration.	46
Tabelle 14-3: Unterstützte Matching-Modes	46
Tabelle 14-4: Unterstützte Notifications im E-PIX.	48
Tabelle 14-5: Elemente der Bloomfilter-Konfiguration.	51
Tabelle 14-6: Unterstützte Algorithmen zur Generierung von Bloomfiltern.	53
Tabelle 14-7: Unterstützte Transformationen für <code>complex-transformation-type</code>	56
Tabelle 14-8: Empfohlene und Standard-Schwellwerte für <i>Automatic Match</i> und <i>Possible Match</i>	57
Tabelle 14-9: Verhalten des E-PIX, je nachdem wie das Element <code>use-cemfim</code> definiert wurde.	58
Tabelle 14-10: Unterstützte Algorithmen für das Matching.	61

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.

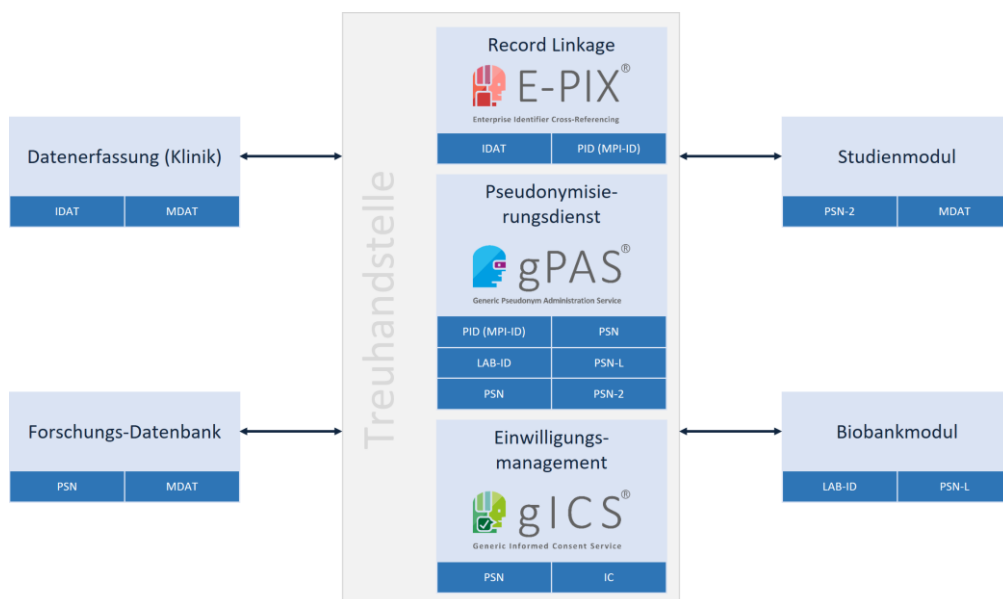


Abbildung 1-1: Das Identitätsdatenmanagement stellt eine zentrale Komponente im medizinischen Forschungskontext dar. Verschiedene Module verwalten modulspezifische Daten und ordnen diese Personen mittels spezifischen Pseudonymen zu. Die Abbildung ist adaptiert vom Maximalmodell des Generischen Datenschutzkonzepts der TMF¹.

Zweck des ID-Managements ist es, Personendaten unter Vermeidung von Homonymfehlern sicher bereits vorhandenen Datensätzen zuzuordnen und potentielle Dubletten zu erkennen und zusammen

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

zu führen. Ergebnis dieser Zuordnung ist eine systemübergreifende eindeutige 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 lizenziert (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-Profiles (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 (kurz: Identität)

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

Referenzidentität / Hauptidentität

Die Referenzidentität ist die Hauptidentität einer Person. Die Hauptidentität gilt als korrekte Ausprägung und wird beispielsweise bei einer Suche in der Ergebnisliste angezeigt. Andere Personenidentitäten (Nebenidentitäten/Ausprägungen der IDAT) der Person können in der Detailansicht eingesehen werden.

Identifizier einer Personenidentität

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

Lokaler Identifizier

Ein Lokaler Identifizier ist ein Identifikator, der durch ein externes System vergeben wurde, wie beispielsweise einem KIS-System. Der Lokale Identifizier identifiziert dabei die Personenidentität eindeutig in diesem System. Aus einem System können dabei mehrere Lokale Identifizier stammen (z.B. Patienten-ID und Fallnummer). Der Personenidentifikator (PID) kann in seiner Funktion als Identifizier auch als LID ("*Lokaler (externer) Identifizier*") betrachtet werden. Im E-PIX können diese Identifizier einer Person zugeordnet werden. Hierzu wird eine Identifizier Domäne (siehe Identifizier Domäne) angelegt, welche die Identifizier z.B. eines externen Systems entspricht und diese Identifizier beinhaltet.

Domäne / Domain / Matching-Domäne

Eine Domain ist eine organisatorische Einheit (Mandant), z.B. eine Studie, ein Projekt oder ein Institut. Im E-PIX können mehrere Domänen angelegt werden, in denen in sich geschlossenen Personenbestände verwaltet werden können. Eine Person ist immer nur innerhalb einer Domäne eindeutig und kann daher in mehreren Domänen auftreten. Das Record Linkage findet nur innerhalb einer Domäne statt. Neben den (Matching-)Domänen, verwaltet der E-PIX Identifizier Domänen (siehe Identifizier Domäne).

Identifizier Domäne

Domäne des Lokalen Identifiziers. Diese muss nicht dem Quellsystem entsprechen. Aus einem Quellsystem können mehrere Lokale Identifizier 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.), welches 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, um derselben 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

Angabe der Datenquelle, aus denen 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. Jeder (Matching-)Domäne kann eine *Sichere Quelle* zugeordnet werden. IDAT die über die *Sichere Quelle* registriert werden, werden als korrekte Ausprägung einer Personenidentität angesehen (Hauptidentität).

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 (z.B.: Müller, Mueller, Muller, Mülller, etc.) im jeweiligen Quellsystem zu erhalten und dennoch die Datensätze eindeutig einer real existierenden Person fehlerfrei zuordnen zu können.

Innerhalb des E-PIX kann eine Person daher mehrere (Personen-)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.

Insbesondere bei der Verwaltung von IDAT, die aus mehreren Quellen stammen, in Abhängigkeit der Eingabemethode und Zeitpunkt der IDAT verschiedene Ausprägungen entstehen (Tippfehler, Namensänderung durch Heirat, etc.). Der E-PIX vereint all diese Ausprägungen zu einer Person und

ermöglicht, die Person über die verschiedenen Ausprägungen zu finden. Mittels der Hauptidentität ist es möglich, die korrekte Ausprägung anzugeben und so bei Bedarf andere Systeme zu aktualisieren.

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. Diese erfolgt im Bedarfsfall vor der Eintragung in den E-PIX.
- 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 Docker² und Docker-Compose³
- Administrative Rechte
- Keine Nutzungsbeschränkungen auf die bereitgestellten Service- und Client-URLs
- 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 zusammenschaltet werden müssen, wird der Dienst mit Docker-Compose gestartet. Die entsprechenden Ressourcen können von der THS-Webseite⁵ heruntergeladen werden.

² Weitere Informationen unter <https://docs.docker.com/install/>

³ Weitere Informationen unter <https://docs.docker.com/compose/install/>

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>

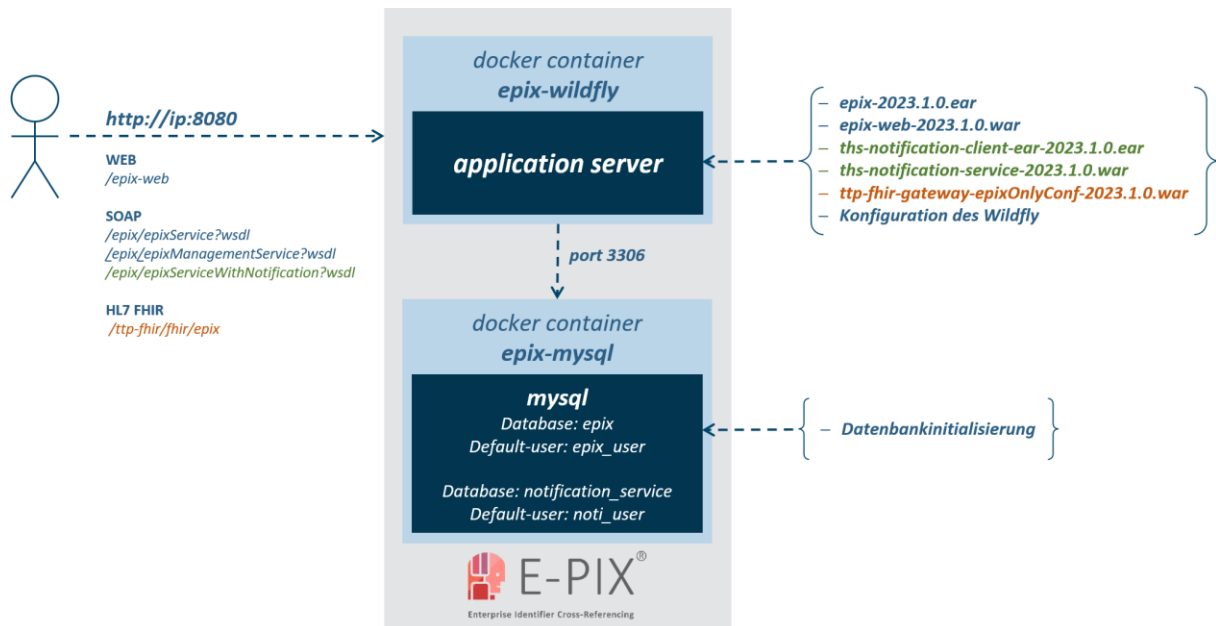


Abbildung 6-1: Architektur des E-PIX® mit Docker.

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 heruntergeladen⁶ 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_E-PIX.md` zu entnehmen.

⚠ Hinweis zum Produktivbetrieb: Für einen Produktivbetrieb sollte die `docker-compose.yml` angepasst werden. Hierzu sollte der Speicherpfad des MySQL-Volumes festgelegt werden. Andernfalls sind alle Daten, die im Container liegen, nach einem Herunterfahren gelöscht. Die Datenbank-Skripte prüfen selbst, ob die entsprechenden Datenbanken bereits angelegt wurden. Die Datenbanken werden bei einem Neustart daher **nicht** überschrieben.

⚠ Hinweis zur Aktualisierung von Containern: Beachten Sie, dass beim Wechsel von E-PIX Versionen die Docker-Compose Komponenten stets komplett aktualisieren sollten. Dies beinhaltet die Aktualisierung von *.yml-Dateien, CLI-Dateien und die Übernahme eventueller individueller Konfigurationen auf neue ENV-Files.

Eine Übersicht aller Konfigurationsdateien, deren Zweck und aller relevanten Parameter ist der beigelegten Beschreibung `docker-compose/README_E-PIX.md` zu entnehmen. Eine ausführliche Anleitung zur Aktualisierung von produktiv genutzten Containern ist dem Produkt beigelegt (`Docker-Update.md`) und online verfügbar (<https://www.ths-greifswald.de/e-pix/update>).

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

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/>).

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

The screenshot shows the 'Einstellungen' (Settings) page in the E-PIX system. The left sidebar contains navigation options: Dashboard, Personen (with sub-options: Dublettenauflösung, Suchen / Bearbeiten, Hinzufügen), Listen (with sub-options: Import, Export), and Administration (with sub-options: Domänen, Protokolle, Info). The 'Aktive Domäne' is set to 'Demo'.

The main content area is divided into three sections, each with a table and a '+ Erstellen' button:

- Domänen verwalten:** A table with columns: Name, Schlüssel, Modus, MPI Identifier-Domäne, and Sichere Datenquelle. It lists 'Demo (aktiv)' and 'Demo-Domäne'.
- Datenquellen verwalten:** A table with columns: Name and Schlüssel. It lists 'Krankenhausinformationssystem' and 'dummy_safe_source'.
- Identifier-Domänen verwalten:** A table with columns: Name, Schlüssel, and OID. It lists 'MPI'.

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 *Erstellen* 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. Es können mehrere Quellen angegeben werden. Pro Domäne wird aber nur eine *Sichere Quelle* angegeben. Bei einer späteren Personenregistrierung kann dann aus der Liste der angelegten Quellen gewählt werden, woher der Datensatz stammt.

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. Im E-PIX ist standardmäßig bereits eine Identifier-Domäne für die *MPIs* angelegt. Diese kann beim Anlegen einer Matching-Domäne als MPI-Domäne angegeben werden. Der E-PIX erzeugt in diese Identifier-Domäne dann die eindeutigen Kennungen, für die später registrierten Personen. Diese Identifier-Domäne kann für mehrere Matching-Domänen eingetragen werden. Dabei werden dann Matching-Domänen übergreifend eindeutige Kennungen vergeben. Soll für jede Matching-Domäne eine eigene MPI-Domäne verwendet werden, so muss für jede Matching-Domäne zunächst eine Identifier-Domäne angelegt werden und bei der Konfiguration als MPI-Domänen angegeben werden. Dabei ist zu beachten, dass *MPIs* im E-PIX immer eindeutig sein müssen. Es ist daher erforderlich, dass in den Matching-Domänen verschiedene Präfixe (siehe **Abschnitt 7.1.3.1**) angegeben werden. Dies ist nicht erforderlich, wenn eine übergreifende MPI-Domäne genutzt wird.

7.1.3 Anlegen einer neuen (Matching-)Domäne

⚠ Hinweis: Ab Version 2023.1.0 kann die Domänenkonfiguration vollständig in der Oberfläche durchgeführt werden. Zuvor war es nötig, die Konfiguration im XML-Format zu hinterlegen. Dies ist weiterhin über die SOAP-Schnittstelle (vgl. **Kapitel 13**) möglich. Eine detaillierte Beschreibung der Konfiguration ist im **Kapitel 14** zu finden.

Nachdem die *Sichere Quelle* und die *Identifier-Domäne* angelegt wurden, kann ein neuer Domänen-Eintrag über die Schaltfläche *Erstellen* erzeugt werden. Die Konfiguration der Domäne erfolgt in mehreren Schritten. Hierfür können verschiedene Reiter ausgewählt werden und die entsprechenden Einstellungen darin vorgenommen werden. Einige Felder sind bereits entsprechend einer Standard-Konfiguration (vgl. **Abschnitt 14.3**) vorausgefüllt, die bei Bedarf angepasst werden können.

⚠ Hinweis: Nach der ersten Personenregistrierung in eine Domäne, kann die Konfiguration nur noch eingeschränkt bearbeitet werden. Andernfalls müsste der E-PIX alle Ergebnisse des Record Linkages anhand der neuen Konfiguration prüfen und ggf. zusammengeführte Identitäten auftrennen. Soll tatsächlich eine neue Konfiguration auf einen Bestand angewandt werden, muss eine neue Domäne angelegt werden und alle Datensätze der vorhandenen Domäne dort registriert werden.

Die Beschreibung der Domänenkonfiguration mit den einzelnen Reitern erfolgt im Folgendem.

7.1.3.1 Einstellungen

Unter dem Reiter Einstellungen werden der Name, die Beschreibung, die *Sichere Quelle*, die *Identifier-Domäne* und weitere allgemeine Einstellungen vorgenommen.

The screenshot shows the 'Einstellungen' (Settings) page for the Enterprise Identifier Cross-Referencing application. The page is organized into several sections:

- Name ***: A text input field.
- Schlüssel**: A text input field.
- Beschreibung**: A text area with a note '255 Zeichen verbleibend'.
- Sichere Datenquelle**: A dropdown menu currently showing 'dummy_safe_source'.
- Sende Benachrichtigungen durch die Weboberfläche**: A checkbox that is currently unchecked.
- Master Patient Index (MPI)**: A section containing:
 - Generator**: A dropdown menu set to 'EAN13Generator'.
 - Präfix ***: A text input field.
 - Identifizier-Domäne**: A dropdown menu set to 'MPI'.
- Geschwindigkeit**: A section containing:
 - Aktiviere paralleles Matching ab**: A text input field set to '1.000' and a button labeled 'Identitäten'.
 - Limitiere Suche auf Matching-Felder (reduziert Speichernutzung aber verhindert Suche nach anderen Feldern)**: A checkbox that is currently unchecked.

Eine Domäne muss einen eindeutigen Namen aufweisen. Der E-PIX erzeugt anhand dessen einen Schlüssel (der wahlweise auch manuell definiert werden kann), welcher zum Ansprechen der Domäne über die SOAP-Schnittstelle verwendet wird. Der Name wird in der Oberfläche angezeigt und kann zu einem späteren Zeitpunkt geändert werden. Der Schlüssel hingegen kann nachträglich nicht mehr geändert werden und bleibt daher beim Ansprechen über die SOAP-Schnittstelle auch nach einer Änderung des Namens unverändert. Eine Beschreibung sollte insbesondere bei der Verarbeitung von Personen für mehrere Mandanten oder Projekte innerhalb eines E-PIX eingetragen werden. Die *Sichere Quelle* kann aus der Liste der vorhandenen Einträge ausgewählt werden. Mit Aktivierung des Kontrollkästchens *Sende Benachrichtigungen...*, benachrichtigt der E-PIX den Notification-Service (vgl. **Kapitel 9**), bei Änderungen in der Oberfläche (z.B. nach Bearbeitung eines Personendatensatzes).

Der E-PIX erzeugt für jede Person einen *MPI*. Der E-PIX wird hierfür mit einem entsprechenden Generator (*EAN13Generator*) ausgeliefert. Soll der *MPI* ein anderes Format aufweisen, können eigene Generatoren implementiert werden. Das Präfix gibt dabei an, ob und welche Zeichenkette einem *MPI* vorangestellt wird (empfohlen: 1001). Der *EAN13Generator* berücksichtigt dieses Präfix, eine etwaige eigene Implementierung muss dies nicht. Zusätzlich wird die Identifizier-Domäne ausgewählt, in der die *MPIs* erzeugt werden sollen (der E-PIX hat standardmäßig hierfür die Identifizier-Domäne „MPI“ hinterlegt).

Zur Verbesserung der Performance können weitere Einstellungen vorgenommen werden. Diese Einstellungen können in der Regel unverändert bleiben. Der E-PIX führt dabei standardmäßig, bevor 1.000 Identitäten registriert wurden, das Record Linkage seriell durch. Danach werden Berechnungen auf einem Mehrkern-System auf die verschiedenen Prozessorkerne aufgeteilt. Zudem kann der Arbeitsspeicherbedarf reduziert werden, indem nur die Felder, die für das Record Linkage erforderlich sind, im Arbeitsspeicher bleiben. Dabei ist zu beachten, dass dabei auch die Suche auf diese Felder beschränkt wird.

7.1.3.2 Personenfelder

Unter dem Reiter Personenfelder werden die Pflichtfelder und Zusatzfelder festgelegt.

Pflichtfelder

Vorname X Nachname X Geschlecht X Geburtsdatum X

Zusatzfelder

Feld	Bezeichnung
Keine Datensätze gefunden.	

+ Zusatzfeld hinzufügen

Standardmäßig sind die Felder Vorname, Nachname, Geschlecht und Geburtsdatum als *Pflichtfelder* hinterlegt. Bei Bedarf kann diese Restriktion durch entfernen der Einträge aufgehoben werden. Dabei ist zu beachten, dass mindestens die Felder, die später für das Record Linkage verwendet werden sollen, als Pflichtfelder anzugeben sind. Pflichtfelder müssen bei einer Personenregistrierung ausgefüllt sein. Weitere Felder können aus der Liste ausgewählt werden.

Darüber hinaus können *Zusatzfelder* definiert werden (die bei Bedarf auch als Pflichtfelder gesetzt werden können). Der E-PIX hat hierfür zehn Freitextfelder, die aus einer Liste gewählt werden können (*Zusatzfeld hinzufügen*). Dabei ist zu beachten, dass diese Felder Restriktionen bzgl. der Länge der eingegebenen Daten aufweisen. Die maximale Anzahl der Zeichen, ist hinter dem jeweiligen Feld angegeben (oder vgl. *value1 - value10* in **Tabelle 14-1**). Für jedes Zusatzfeld kann ein Bezeichner gewählt werden, der bei der Personenregistrierung am entsprechenden Feld steht.

7.1.3.3 Vorverarbeitung

Bei der Personenregistrierung eingegebene IDAT können für ein Record Linkage aufbereitet werden. Dies umfasst bspw. das Entfernen von unerwünschten Zeichenketten oder die Vereinheitlichung von Umlauten. Dies betrifft aber nur die interne Verarbeitung. Die IDAT werden wie eingegeben in der Oberfläche dargestellt. Die Vorverarbeitung verbessert das Record Linkage und damit die Zusammenführung von Datensätzen, die zu einer Person zugehörig sind.

Felder mit Vorverarbeitungen

Feld	Ersetzungen	Umwandlungen	Filter	
Vorname	23	2	0	
Nachname	23	2	0	

+ Feld zur Vorverarbeitung hinzufügen

Standardmäßig sind für die Felder Vorname und Nachname entsprechende Vorverarbeitungen hinterlegt. Diese können bearbeitet oder entfernt werden. Zudem kann für weitere Felder eine Vorverarbeitung definiert werden. Der E-PIX unterscheidet zwischen Ersetzungen, Umwandlungen und Filtern. Es können jeweils mehrere Vorverarbeitungen pro Feld hinterlegt werden. Bei einer Ersetzung wird eine definierte Zeichenkette, mit einer anderen ersetzt (wenn die ersetzende Zeichenkette leer ist, wird die zu ersetzende Zeichenkette entfernt. Bsp.: Zu ersetzen: „Dr.“, „Ersetzung: „“. Damit wird die Zeichenkette „Dr.“ restlos aus dem entsprechenden Feld entfernt.). Dabei ist zu beachten, dass die Groß- und Kleinschreibung berücksichtigt wird. Für Standardfälle, wie die Ersetzung von Umlauten, gibt es Umwandlungen. Der E-PIX wird mit vier Umwandlungen ausgeliefert:

ToUpperCaseTransformation: Ersetzt alle Zeichen durch den entsprechenden Großbuchstaben. Beim Record Linkage werden so Unterschiede bei der Groß- und Kleinschreibung nicht berücksichtigt.

CharsMutationTransformation: Ersetzt alle Umlaute: „ä“ durch „ae“, „Ä“ durch „AE“, „ü“ durch „ue“, „Ü“ durch „UE“, „ö“ durch „oe“, „Ö“ durch „OE“ und „ß“ durch „SS“.

CharNormalizationTransformation: Überführt eine Zeichenkette in ASCII⁷. Dies entfernt z.B. Akzente. Dabei ist zu beachten, dass Umlaute wie ä nicht in ae, sondern in a überführt. Eine Kombination mit *CharsMutationTransformation* ist möglich.

TrimTransformation: Entfernt führende und folgende Leerzeichen. Bsp.: „ Müller “ → „Müller“.

Beim Filtern kann ein Alphabet mit zulässigen Zeichen angegeben werden. Alle anderen Zeichen, werden bei der Vorverarbeitung durch das angegebene Zeichen ersetzt. Wenn letzteres leer ist, dann werden unzulässige Zeichen entfernt. Dieser Filter sollte nur dann angewandt werden, wenn die Menge der zulässigen Zeichen bekannt ist (z.B. die Postleitzahl darf nur Zahlen enthalten) oder begrenzt werden muss (z.B. um Bloomfilter zu erzeugen).

7.1.3.4 Matching

Unter dem Reiter Matching werden die Parameter für das Record Linkage festgelegt. Dies umfasst das Setzen von Schwellwerten, also ab wann zwei Datensätze zu einer Person zugeordnet werden und welche Felder für den Abgleich verwendet werden sollen.

Matching

Führe Record Linkage durch Ja Nein

Prüfe, ob übergebene identifizierende Daten neben der Person des Identifiers gleich gut mit anderen Personen matchen. [?](#)

Mindestscore für möglichen Match * häufig 2.99 selten

Mindestscore für automatischen Match * häufig 14.5 selten nie

Matching Felder

Feld	Algorithmus	Modus	Blocking ab Übereinstimmung von	Matching ab Übereinstimmung von	Gewichtung	
Vorname	LevenshteinAlgorithm	Text	40 %	80 %	8.0	
Nachname	LevenshteinAlgorithm	Text	0 %	80 %	6.0	
Geschlecht	LevenshteinAlgorithm	Text	0 %	75 %	3.0	
Geburtsdatum	LevenshteinAlgorithm	Zahlen	60 %	100 %	9.0	

+ Matching Feld hinzufügen

Der E-PIX unterscheidet zwischen zwei Modi. Zum einen kann der E-PIX Personendaten selbst mittels Record Linkage zusammenführen, *MPIs* vergeben usw. (Führe Record Linkage durch: ja). Es besteht ebenso die Möglichkeit, dass der E-PIX Personendatensätze nur ablegt (Führe Record Linkage durch: nein). Dies kann gewünscht sein, wenn ein Record Linkage bereits in einem anderen System durchgeführt wurde (z.B. in einem KAS). In beiden Fällen wird eine Matching-Konfiguration hinterlegt, damit der E-PIX Personendatensätze korrekt zuordnen kann. Sollen die Personendatensätze nur

⁷ https://de.wikipedia.org/wiki/American_Standard_Code_for_Information_Interchange

abgelegt werden, erfolgt dies unter bestimmten Bedingungen. Beispielsweise müssen zwei Personendatensätze mit derselben übergebenen (externen ID) komplett übereinstimmen, oder zumindest laut der angegebenen Konfiguration eine gewisse Übereinstimmung aufweisen. Soll der E-PIX ein Record Linkage durchführen, bestimmt die Konfiguration, wann zwei Personendatensätze zur selben Person als Identitäten zugeordnet werden und dementsprechend dieselbe *MPI* erhalten.

Beim Record Linkage klassifiziert der E-PIX die Datensätze in Match-Typen (vgl. **Kapitel 3**). Ein möglicher Match (Possible Match) entsteht, wenn die Übereinstimmung über dem Schwellwert für einen möglichen Match liegt, jedoch niedriger als der Schwellwert für einen automatischen Match ist. Bei einem möglichen Match kann später manuell entschieden werden (siehe **Abschnitt 7.6**), ob zwei Datensätze zur selben Person zugehörig sind, oder zwei verschiedene Personen darstellen. Bei der Entscheidung können entsprechend weitere Informationen zugezogen werden. Sind beide Schwellwerte identisch, so werden keine possible Matches angelegt.

Liegt die ermittelte Übereinstimmung über dem Schwellwert für einen automatischen Match, so führt der E-PIX die entsprechenden Datensätze entsprechend zusammen, auch wenn keine vollständige Übereinstimmung (z.B. durch Tippfehler) vorliegt. Im Ergebnis werden die Datensätze als Identitäten einer Person zugeordnet. Ein automatisches Zusammenführen kann unterbunden werden, indem das Kontrollkästchen „*nie*“ angewählt oder der Wert auf 1000 gesetzt wird.

Die Übereinstimmung zweier Datensätze, ermittelt der E-PIX anhand der definierten *Matching Felder*. Für jedes Feld kann ein Vergleichsalgorithmus, eine Wichtung und Schwellwerte für das Blocking und den Abgleich definiert werden. Der E-PIX unterstützt verschiedene Vergleichsalgorithmen. Für die meisten Fälle ist jedoch der Algorithmus *LevenshteinAlgorithm* zu empfehlen. Dieser ermittelt die Levenshtein-Distanz zweier Zeichenketten, anhand derer die Übereinstimmung berechnet werden kann. Alle unterstützten Algorithmen sind in **Tabelle 14-10** aufgelistet.



Das Blocking beschleunigt das Record Linkage, indem es zunächst nur grob Datensätze miteinander abgleicht und bei hinreichender Übereinstimmung alle *Matching Felder* zum Abgleich verwendet. Der Schwellwert sollte daher nicht zu hoch gewählt werden, damit das Blocking nicht Datensätze aussortiert, die bei einem genaueren Vergleich einer Person zugeordnet werden würden. Der Modus gibt an, welcher Datentyp im Feld enthalten ist (Text oder Zahlen) und betrifft nur das Blocking. Dieser optimiert den internen Abgleich und wird in den meisten Fällen auf „Text“ gesetzt.

Der Schwellwert für das Matching gibt an, ab welcher Übereinstimmung zwei *Matching Felder* übereinstimmen. Das Ergebnis fließt der angegebenen *Gewichtung* entsprechend, in das Ergebnis mit ein. Dabei ist zu beachten, dass wenn es nur ein Matching-Feld gibt, muss der Wert auf 100 gesetzt werden. Wird anhand aller *Matching Felder* eine der oben genannten Schwellwerte überschritten, werden die betreffenden Datensätze entsprechend als possible Match oder automatischen Match klassifiziert. Andernfalls wird der zu registrierende Datensatz als *No Match* klassifiziert und entsprechend als neue Person angelegt.

Felder können als *Multi-Wert Feld* angegeben werden. Dabei werden die Inhalte eines Feldes anhand eines *Trennsymbols* aufgeteilt und separat abgeglichen. Wird z.B. erwartet, dass im Feld Vorname mehrere Vornamen angegeben werden, können so die einzelnen Vornamen zwischen zwei Personendatensätzen abgeglichen werden. Eine detailliertere Beschreibung, inkl. der hierfür anzugebenden Schwellwerte, ist in **Kapitel 14.4.12.7** zu finden.

⚠ Hinweis: Die Konfiguration basiert auf Erfahrungswerten und ist häufig projektabhängig. Je nach zu erwartender Datenqualität können höhere Schwellwerte gewählt werden, um beispielsweise weniger possible Matches zu erzeugen.


Bei der Dublettenauflösung können Gründe angegeben werden. Dies erfolgt mittels Freitextfeld. Für häufig auftretende Gründe, können entsprechende Vorlagen definiert werden.

Gründe für Dublettenauflösung		
Bezeichnung	Hinweis	
Tippfehler	Vertauschte, fehlende oder zu viele Zeichen	
Namensänderung durch Heirat	Änderung des Nachnamens auf Grund einer Heirat	
+ Grund hinzufügen		

Hierfür wird für jeden Grund ein Bezeichner gewählt, der bei der Dublettenauflösung angewählt werden kann. Der angegebene Hinweis wird dann entsprechend protokolliert.

7.1.3.5 Privatsphäre

Der E-PIX ermöglicht das Anlegen von Bloomfiltern, um ein Privacy-Preserving Record Linkage durchzuführen. Dies kommt normalerweise bei standortübergreifenden Abgleichen zum Einsatz. Der E-PIX kann sowohl Bloomfilter anlegen, als auch miteinander vergleichen. Der Vergleich wird mittels *Matching Felder* definiert. Standort-interne Vergleiche finden üblicherweise über die Klartextdaten statt. Standardmäßig wird kein Bloomfilter angelegt. Die Konfiguration erfolgt üblicherweise projektspezifisch.

Einstellungen	Personenfelder	Vorverarbeitung	Matching	Privatsphäre
Speichermodus Identifizierende Daten und Bloomfilter				
Bloomfilter				
Algorithmus	Quell-Felder		Speicherfeld	
RandomHashingStrategy	Vorname Nachname Geburtsdatum Geschlecht		Feld 6	 
+ Bloomfilter hinzufügen				

Im Bild wurde exemplarisch die Konfiguration eines Bloomfilters hinterlegt.

⚠ Hinweis: Der E-PIX unterstützt mehrere Algorithmen zur Erzeugung von Bloomfiltern 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.

Über die *Schaltfläche Bloomfilter hinzufügen* wird eine neue Bloomfilter-Konfiguration angelegt. Zunächst wird der zu verwendende Algorithmus angegeben. Eine Auflistung mit kurzer Erläuterung ist in **Tabelle 14-6** zu finden.

Je nach verwendetem Algorithmus, kann ein Alphabet angegeben werden. Dabei ist zu beachten, dass zur Bloomfilter-Generierung die vorverarbeiteten Werte verwendet werden. Damit muss sichergestellt werden, dass die verwendeten IDAT-Felder so vorverarbeitet (siehe **Abschnitt 7.1.3.3**) wurden, dass diese nur Zeichen enthalten, die auch im angegebenen Alphabet enthalten sind. Besteht das Alphabet

nur aus Großbuchstaben, so sollte zuvor das Feld zuvor mit *ToUpperCaseTransformation* transformiert worden sein. Umlaute sollten zuvor mit *CharsMutationTransformation* und Akzente etc. per *CharNormalizationTransformation* entfernt worden sein. Mit einem Filter kann sichergestellt werden, dass Felder nur Zeichen beinhalten, die auch im Alphabet vorkommen. Zu beachten ist, dass die Groß- und Kleinschreibung beachtet wird. Sollen die Zustände vom Feld Geschlecht berücksichtigt werden (intern kodiert mit m, f, o, u, x), so müssen diese Zeichen entsprechend auch im Alphabet vorkommen.

Die Länge gibt die Anzahl der Bits pro Bloomfilter an. Zwar ist die Wahl des Speicherfeldes frei, jedoch ist zu beachten, dass der E-PIX die Feldlängen intern begrenzt. Außerdem werden die Bloomfilter intern im Base64-Format kodiert. Die meisten Felder vom E-PIX erlauben eine maximale Länge von 255 Zeichen⁸. Werden längere Bloomfilter benötigt, sollten die frei definierbaren Felder (*value8* - *value10*, vgl. **Tabelle 14-1**) verwendet werden. Die tatsächlich benötigte Länge kann durch die Verwendung von Härtungsverfahren beeinflusst werden. So halbiert jede Faltung beim *XOR-Folding* die resultierende Länge. Die Nutzung eines *Balanced Bloomfilters* verdoppelt die resultierende Länge.

Mit der Länge der N-Gramme wird angegeben, wie lang die Teil-Zeichenketten beim kodieren der Felder in den Bloomfilter sein sollen. Üblicherweise werden hierfür Bigramme (N=2) genutzt.

Mit Bits pro N-Gramm kann die Anzahl der Bit-Positionen pro N-Gramm angegeben werden. Je höher dieser Wert gewählt wird, desto mehr Positionen werden im resultierenden Bloomfilter belegt.

Die Anzahl der XOR-Faltungen (*XOR-Folding*⁹) gibt an, wie oft ein Bloomfilter gefaltet werden soll. Dies härtet den Bloomfilter gegen Angriffe. Mit jeder Faltung halbiert sich die Länge des Bloomfilters. Zu beachten ist, dass die Anzahl der Faltungen ein ganzzahliger Teiler der Länge sein muss. Die Anzahl der Faltungen sollte gering gehalten werden, da andernfalls die Qualität des Record Linkages negativ beeinflusst werden kann.

Mit der Aktivierung des Kontrollkastens *Balanced Bloomfilter*¹⁰, wird bei der Erzeugung des Bloomfilters eine negierte Kopie angefügt und die Bit-Positionen mittels des angegebenen Werts (*Seed*) zufällig vertauscht. Der *Seed* muss eine Ganzzahl sein.

Das *Speicherfeld* gibt an, in welchem Feld der resultierende Bloomfilter gespeichert werden soll. Dabei muss beachtet werden, dass zum einen der Bloomfilter in das ausgewählte Feld passt (siehe auch *Länge*) und zum anderen, dass etwaige Informationen im Feld überschrieben werden (Bsp.: Wenn das Feld Vorname als Speicherfeld gewählt wurde, ist nach einer Personenregistrierung der Vorname durch den Bloomfilter überschrieben). Es ist daher ratsam, dass Speicherfeld auf ein Value-Feld (*Zusatzfeld*) zu setzen.

⁸ Um die benötigte Speicherlänge zu ermitteln, kann folgende Formel angewandt werden: $z = 4 \times \left\lceil \frac{\text{Bits}}{8} \right\rceil$

Für eine Länge von 1000 Bits ergibt sich ein Bedarf von $4 \times \left\lceil \frac{1000}{8} \right\rceil = 167$ Zeichen.

⁹ Schnell, Rainer and Borgs, Christian, XOR-Folding for Bloom Filter-based Encryptions for Privacy-preserving Record Linkage (December 22, 2016). German Record Linkage Center, NO. WP-GRLC-2016-03, DECEMBER 22, 2016, Available at SSRN: <https://ssrn.com/abstract=3527984> or <http://dx.doi.org/10.2139/ssrn.3527984>

¹⁰ R. Schnell and C. Borgs, "Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 218-224, doi: 10.1109/ICDMW.2016.0038.

Jedem Bloomfilter können beliebig viele *Quell-Felder* zugeordnet werden. Auf Basis der darin enthaltenen Werte, wird bei der Registrierung der Bloomfilter erzeugt. Je nach Verfahren muss zusätzlich ein *Seed* (als Ganzzahl), ein *fester Salt* (beliebige Zeichenkette) oder ein Feld als *Salt* angegeben werden. Ein *Salt* ist ein Wert, der intern vor der Kodierung jedem N-Gramm angefügt wird. Wird ein Feld als *Salt* gewählt, so wird vom jeweiligen Datensatz der Wert des Feldes hierzu verwendet. Hierzu eignen sich festgelegte Pflichtfelder (z.B. das Geburtsdatum).

❶ Field-Level oder Record-Level Bloomfilter?

Der E-PIX unterstützt sowohl die Erzeugung von Field-Level Bloomfiltern (ein Bloomfilter pro Feld), als auch die Erzeugung von Record-Level Bloomfiltern (Bloomfilter kodiert mehrere Felder). Zum Erzeugen von Field-Level Bloomfiltern, wird pro Feld ein Bloomfilter definiert. Dabei wird als Quell-Feld nur das entsprechende Feld ausgewählt. Beim Record-Level Bloomfilter werden mehrere Quell-Felder angegeben, die alle im selben Bloomfilter kodiert werden.

Soll der E-PIX nur zur Erzeugung von Bloomfiltern genutzt werden (bspw. weil die Verwaltung der IDAT in einem anderen System erfolgt), so kann der *Speichermodus* zu *Nur Bloomfilter* geändert werden. Die angegebenen Quell-Felder werden nur zu Generierung des Bloomfilters verwendet. Alle IDAT-Felder werden nicht persistiert. Ein Record Linkage kann dann nur über die Bloomfilter durchgeführt werden. Standardmäßig werden sowohl Bloomfilter, als auch IDAT-Felder persistiert.

7.2 Registrierung 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 Auswahlmü 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 14.4.7**). 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 14.4.8**). 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.

⚠ Hinweis: Jeder Domäne wird eine MPI-Domäne zugeordnet. In diese MPI-Domäne erzeugt der E-PIX die MPIs. Diese MPI-Domäne kann deshalb nicht für Externe/Lokale Identifier ausgewählt werden. Hierzu muss zunächst eine entsprechende Domäne angelegt werden (siehe **Abschnitt 7.1.2**), die den Bereich der Identifier darstellt (z.B. Fallnummern, ID im KAS, usw.).

❗ 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 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.

The screenshot shows the 'Hinzufügen' (Add) form in the E-PIX system. The form is divided into several sections:

- Stammdaten (Personal Data):** Includes fields for Title, First Name (Vorname *), Last Name (Nachname *), Gender (Geschlecht *), Birth Date (Geburtsdatum *), Birth Place (Geburtsort), Birth Name (Geburtsname), and Vital Status (Vitalstatus).
- Projektdaten (Project Data):** Includes Nationality & Religion (Nationalität, Ethnische Zugehörigkeit, Religion, Muttersprache) and Other Data (Sonstige Daten) such as Middle Name (Mittelnamen), Family Status (Familienstand), Prefix, Suffix, and External Date (Externes Datum).
- Adresse (Address):** Includes Street and Number (Straße und Nr.), Postal Code (PLZ), City (Stadt), District (Kreis), and State (Bundesland).
- Lokale Identifier (Local Identifier):** A section with a red plus sign and the text '+ Lokalen Identifier eingeben'.

The left sidebar contains navigation options: Dashboard, Personen (with sub-options for Duplicates, Search/Edit, and Add), Listen (Import/Export), Administration (Domains, Protocols, Info), and Active Domain (Demo-Domäne).

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.
<i>Good/Automatic Match</i> <i>bzw. Match</i>	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 <i>möglichen Match</i> 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 Suche 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.

The screenshot shows the E-PIX search interface. The top bar includes the E-PIX logo and the title 'Suchen / Bearbeiten'. A navigation sidebar on the left contains sections for 'Dashboard', 'Personen' (with sub-items: Dublettenauflösung, Suchen / Bearbeiten, Hinzufügen), 'Listen' (with sub-items: Import, Export), 'Administration' (with sub-items: Domänen, Protokolle, Info), and 'Aktive Domäne' (set to Demo-Domäne). The main content area features an information banner: 'Information: 1 Person zu den eingegebenen identifizierenden Daten gefunden.' Below this is a table with columns: Aktionen, MPI, Vorname, Nachname, Geschlecht, Geburtsdatum, Geburtsort, Vitalstatus, and jH. The first row shows an 'Öffnen' button, MPI '1001000000073', and personal data for Hannah Gasser. Below the table are several form sections: 'MPI' (empty input), 'Lokaler Identifier' (with a dropdown for 'Bitte wählen' and an identifier input), 'Stammdaten' (with fields for Titel, Vorname 'Hannah', Nachname 'Gasser', and Geschlecht 'Weiblich'), and 'Projektdaten' (with fields for Nationalität & Religion, Ethnische Zugehörigkeit, Religion, and Muttersprache).

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

7.4 Einsehen von Details zu einer Person

The screenshot shows the E-PIX details page for a person. The top bar includes the E-PIX logo and the title 'Details zur Person'. The navigation sidebar is similar to the search page but includes 'Suchen / Bearbeiten' under 'Personen'. The main content area is divided into several sections:

- Person Summary:** 'Hannah Gasser', 'Weiblich, 37 Jahre', 'Geboren 11.09.1985 in Memmingen'. Includes a 'Person löschen' button.
- Master Identifier (MPI):** '- 1001000000059' with a 'gPAS' icon.
- Externe Identifier:** 'Dem E-PIX sind bisher keine Identifier für diese Person in externen Systemen bekannt.' Includes a '+ Externen Identifier hinzufügen' button.
- Identitäten:** 'Diese Schreibweisen sind dem E-PIX zur Person bekannt. Die bevorzugte Schreibweise wird als **Referenz** bezeichnet und kann festgelegt werden. Sie können weitere Schreibweisen mit **Identität hinzufügen** angeben.' Below is a table with columns: Aktionen, Status, Vorname, Nachname, Geschlecht, Geburtsdatum, Geburtsort, Vitalstatus, and Blo. The first row shows a 'Referenz' status and the name 'Hannah Gasser'. Includes a '+ Identität hinzufügen' button.
- Adressen:** 'Diese Wohnorte sind dem E-PIX zur Person bekannt.' Below is a table with columns: Aktionen, Status, Straße und Nr., PLZ, Stadt, and Bundesland. The first row shows 'Aktuell' status and the address 'Weiherstraße 34, 32425, Minden, Bayern'.

Abbildung 7-4: Detailseite zur Einsicht von den Stammdaten 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. So lassen sich der Person weitere Identitäten/Ausprägungen oder Adressen hinzufügen. Sind mehrere Identitäten zur Person bekannt, so kann im Abschnitt *Identitäten* die *Hauptidentität* (siehe **Begriffsbestimmungen**) mit der Wahl des Sterns ausgewählt werden. Einzelne Identitäten können in dieser Liste mit der entsprechenden Aktion bearbeitet oder entfernt werden. Existiert zu einer Person nur eine Identität, so ist diese automatisch die Hauptidentität und kann nicht gelöscht werden. Soll der gesamte Personendatensatz aus dem E-PIX entfernt werden, kann im oberen Teil die Schaltfläche *Person löschen* gewählt werden. In **Abbildung 7-4** ist exemplarisch die Detailseite einer Person dargestellt.

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.

The screenshot shows the 'E-PIX' software interface for 'Details zur Person'. A modal window titled 'Identität hinzufügen' is open, containing several data entry sections. The 'Stammdaten' section includes fields for Title, Vorname (Hannah), Nachname (Gässer), and Geschlecht (Weiblich). The 'Geburt & Tod' section includes Geburtsdatum (11.09.1985), Geburtsort (Memmingen), and Vitalstatus (Lebendig). The 'Projektdateien' section includes Nationalität & Religion, Ethnische Zugehörigkeit, Religion, and Muttersprache. The 'Sonstige Daten' section includes Mittelname, Familienstand, Prefix, Suffix, and Externes Datum. At the bottom of the form, there are dropdown menus for 'Grund der Bearbeitung' (Schreibfehler im Nachnamen) and 'Datenquelle' (Krankenhausinformationssystem), a checkbox for 'Neue Identität erzwingen', and buttons for 'Speichern' and 'Abbrechen'. The background shows a sidebar with navigation options like 'Dashboard', 'Personen', 'Dubletten', 'Suchen', 'Import', 'Export', 'Administrativ', 'Domänen', 'Info', 'Aktive Domänen', and 'Demo-Domänen'. A status bar at the bottom of the window displays 'Person und Identität angelegt' with a green plus icon, and a timestamp '19.10.2022 08:18:10' and address 'Weiherstraße 34, 32425 Minden'.

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 neueste 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 Dublettenauflösung

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 ausgewählt. Beide Personendatensätze werden tabellarisch gegenübergestellt und Unterschiede bei den jeweiligen Attributen farbig hervorgehoben (siehe **Abbildung 7-666**). 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 14.4.9** oder **Abschnitt 7.1.3.4**) definiert werden und sind dann wählbar. Dies reduziert bei häufig auftretenden Fehlern die Schreibaarkeit.

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.

The screenshot shows the 'Dublettenauflösung' (Duplicate Resolution) interface in E-PIX. The main area displays a comparison between two persons, 'Person 1' and 'Person 2'. The data is presented in a table format with the following columns: Aufgetreten, Vorname, Nachname, Geburtsdatum, and MPI. Below the table, there are action buttons: 'Zusammenführen zur Person 1', 'Trennen', 'Zurückstellen', and 'Zusammenführen zur Person 2'. The interface also includes a sidebar with navigation options like 'Dashboard', 'Personen', 'Dublettenauflösung', 'Suchen / Bearbeiten', 'Hinzufügen', 'Listen', 'Import', 'Export', 'Administration', 'Domänen', 'Protokolle', 'Info', and 'Aktive Domäne'.

	Person 1				Person 2				
	Aufgetreten	Vorname	Nachname	Geburtsdatum	MPI	Vorname	Nachname	Geburtsdatum	MPI
	18.10.2022 15:49:49	Hannah	Gasser	11.09.1985	1001000000073	Anna	Gässer	11.09.1985	1001000000080
Aktion	Zusammenführen zur Person 1				Zusammenführen zur Person 2				
MPI	1001000000073				1001000000080				
Letzte Änderung	18.10.2022 14:01:25				18.10.2022 15:49:49				
Vorname	Hannah				Anna				
Nachname	Gasser				Gässer				
Geschlecht	Weiblich				Weiblich				
Geburtsdatum	11.09.1985				11.09.1985				
Geburtsort	Memmingen								
Straße und Nr.	Weiherstraße 34								
PLZ	32425								
Stadt	Minden								
Bundesland	Bayern								
Bloomfilter NUM-Projekt A	jHtoGGoaBjwNp0CoLYuP2Xzqp9+SYVNj2hJ [...]				JFtoECgaJngN40DpnYqL+aC6p80XIVPjoyIJ [...]				
Bloomfilter Mil-Projekt X	51d/nqj31mMR7hX5cbmnD9ZCvzDEsn/qQ7Aw [...]				4Vanmqj3GmM87BXSYZmnPjJiv7DE8lqR7B4 [...]				
Anzahl Adressen	1				0				

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 7-777** wird die entsprechende Oberfläche exemplarisch dargestellt.

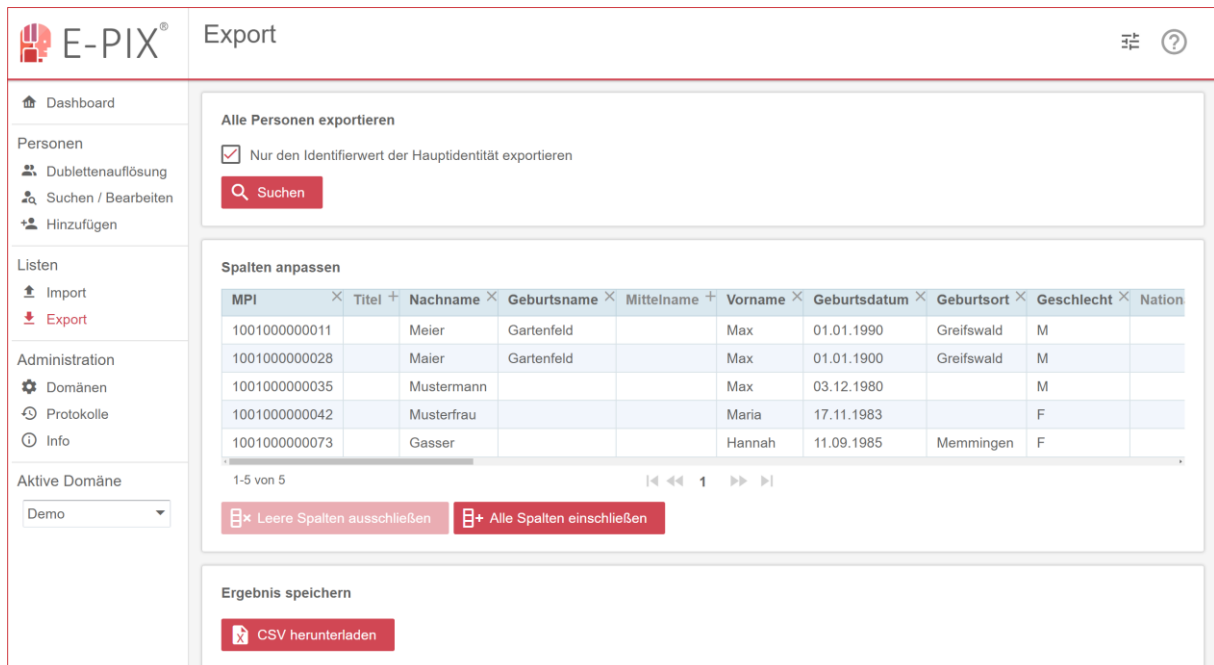


Abbildung 7-7: Oberfläche zum Exportieren von Personendaten.

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.

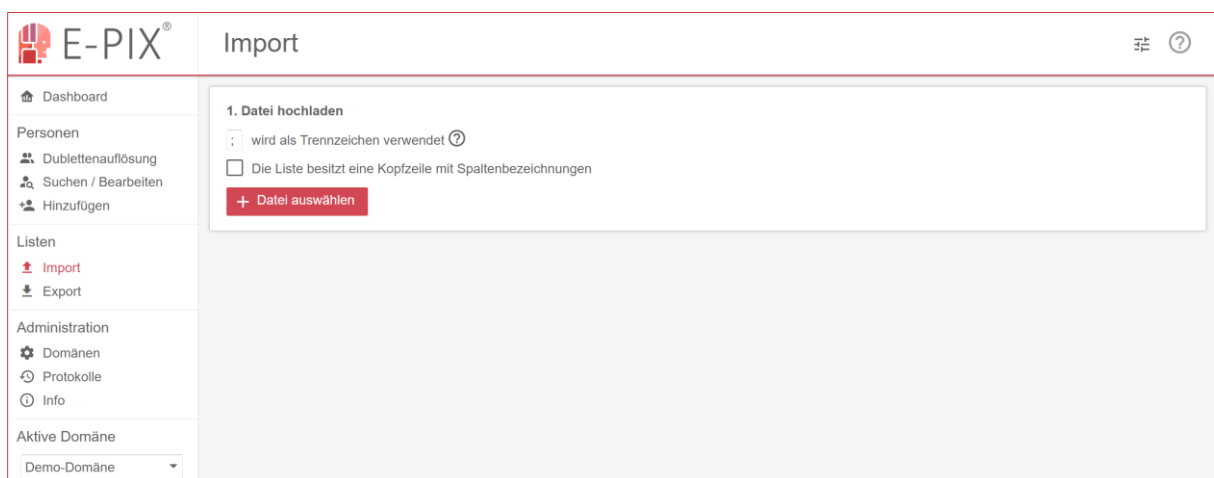
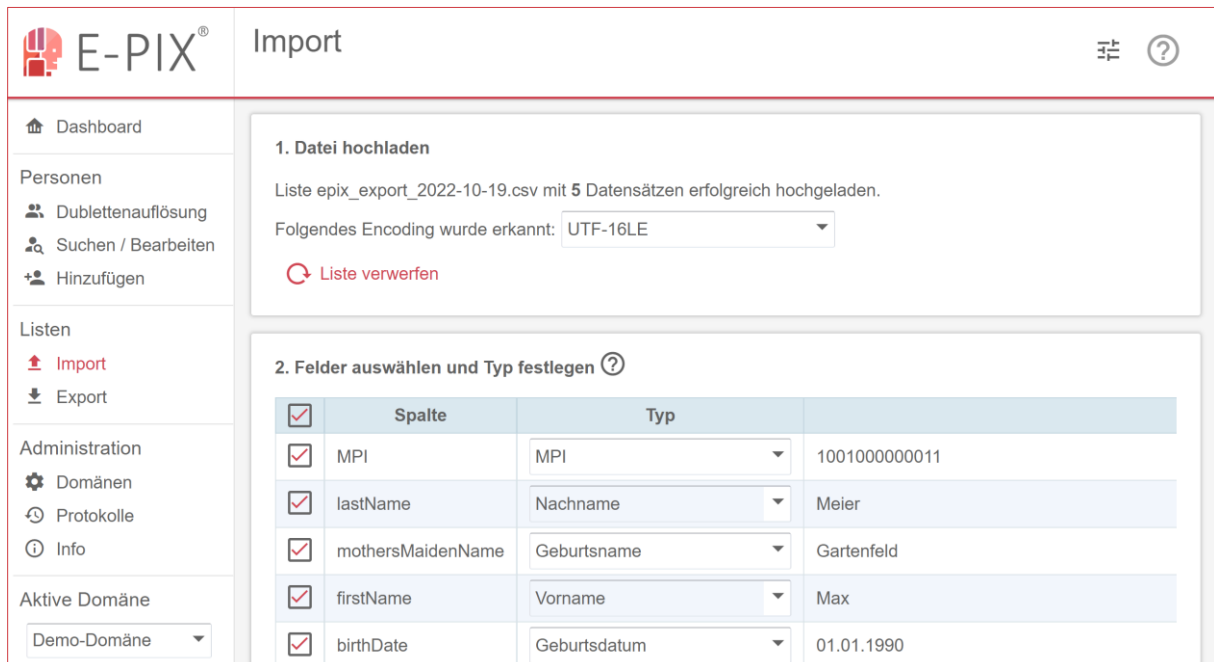


Abbildung 7-8: Oberfläche zum Importieren von Personendaten.

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 werden¹¹.

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.



The screenshot shows the E-PIX 'Import' interface. It features a sidebar with navigation options like 'Dashboard', 'Personen', 'Listen', and 'Administration'. The main content area is divided into two sections:

1. Datei hochladen
 Liste epix_export_2022-10-19.csv mit 5 Datensätzen erfolgreich hochgeladen.
 Folgendes Encoding wurde erkannt: UTF-16LE
 [Liste verwerfen]

2. Felder auswählen und Typ festlegen

<input checked="" type="checkbox"/>	Spalte	Typ	
<input checked="" type="checkbox"/>	MPI	MPI	1001000000011
<input checked="" type="checkbox"/>	lastName	Nachname	Meier
<input checked="" type="checkbox"/>	mothersMaidenName	Geburtsname	Gartenfeld
<input checked="" type="checkbox"/>	firstName	Vorname	Max
<input checked="" type="checkbox"/>	birthDate	Geburtsdatum	01.01.1990

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.

¹¹ Dieser Eintrag wird beim Import nicht als Zeile eingelesen und beeinflusst nicht eine etwaig vorhandene Kopfzeile.

- **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 weiß 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.

The screenshot shows the 'Protokolle' (Records) section of the E-PIX application. It features a sidebar with navigation options like 'Dashboard', 'Personen', 'Listen', and 'Administration'. The main area displays a table titled 'Identitäten Ereignisprotokoll' with columns for 'Zeitpunkt', 'MPI', 'Ereignis', 'Vorname', 'Nachname', 'Geburtsdatum', and 'Geschlecht'. The table contains five rows of data for various individuals. Below the table, there is a 'Download' section with a 'CSV herunterladen' button.

Zeitpunkt	MPI	Ereignis	Vorname (neu)	Nachname (neu)	Geburtsdatum (neu)	Geschlecht (neu)	Vorname (alt)	Nachname (alt)	Ge
19.10.2022 11:13:50	1001000000103	NEW	Maria	Musterfrau	17.11.1983	Weiblich			
19.10.2022 11:13:50	1001000000097	NEW	Max	Mustermann	03.12.1980	Männlich			
19.10.2022 11:13:50	1001000000080	NEW	Max	Maier	01.01.1900	Männlich			
19.10.2022 11:13:50	1001000000073	NEW	Max	Meier	01.01.1990	Männlich			
19.10.2022 09:18:10	1001000000066	NEW	Hannah	Gasser	11.09.1985	Weiblich			

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.

The screenshot shows the 'Dashboard' section of the E-PIX application. It features a sidebar with navigation options like 'Dashboard', 'Personen', 'Listen', and 'Administration'. The main area displays several statistical widgets:

- Statistik für:** Demo-Domäne, Alle Domänen
- Kennzahlen:** 10 Personen, 10 Identitäten, 4 Offene Dubletten, 0 Zusammengeführte Dubletten, 2 Getrennte Dubletten, 18 ms Berechnung, 12:05:33 Datenstand, Aktualisieren
- Personen und Identitäten - Verlauf:** Line chart showing the trend of persons and identities over time.
- Personen und Identitäten - Zuwachs:** Bar chart showing the monthly and yearly growth of persons and identities.
- Personen und Identitäten:** Summary statistics: 1.0 Identitäten pro Person, 0.6 Mögliche Matches pro neuer Identität, 0.0 Automatische Matches pro neuer Identität.
- Matching - Verteilung:** Pie chart showing the distribution of matches: Kein Match, Möglicher Match, Guter Match, Perfekter Match.
- Matching - Verlauf:** Line chart showing the trend of different match types over time.

Abbildung 7-11: Dashboard zum Einsehen der Statistiken.

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

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

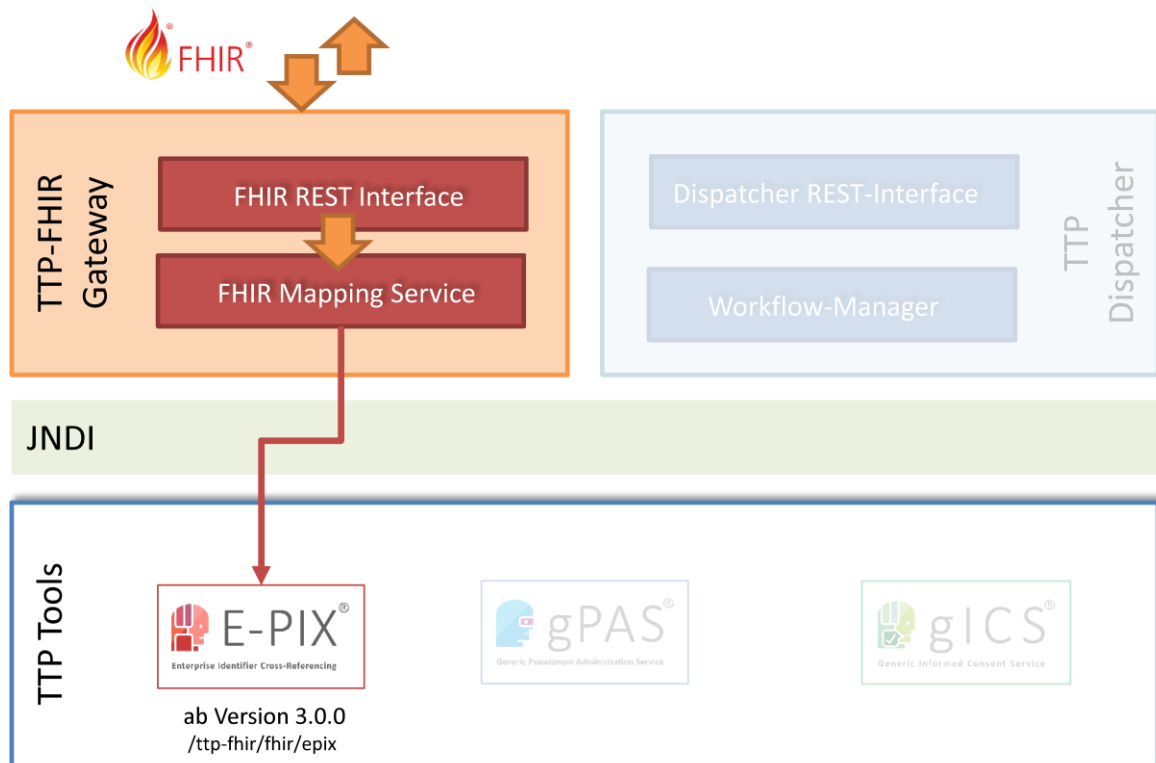
„Fast Healthcare Interoperability Resources (kurz: FHIR®) ist ein von HL7 erarbeiteter Standard. Dieser unterstützt den Datenaustausch zwischen Softwaresystemen im Gesundheitswesen. FHIR beschreibt Datenformate und Elemente als sogenannte „Ressourcen“ und bietet eine Schnittstelle an, um diese auszutauschen“¹².

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

¹² https://de.wikipedia.org/wiki/Fast_Healthcare_Interoperability_Resources, Zugriff am 22.01.2021

Treuhandstellen-FHIR-Gateway (kurz: TTP-FHIR Gateway) als Mittler von FHIR-spezifischen Infrastrukturkomponenten und E-PIX bereitgestellt.

⚠ 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.



© Independent Trusted Third Party Greifswald 2022

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. Gefyra¹³.

11 Authentifizierung und Autorisierung

11.1 Global

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` entnommen werden.

⚠ 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 detaillierte Beschreibung ist unter <https://www.ths-greifswald.de/ttp-tools/keycloak> verfügbar.

¹³ <https://www.gefyra.de/>, Zugriff am 2021-06-08

11.1.1 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			×

11.1.2 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.

11.1.3 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.


11.2 Domänen-spezifische Rollen mit OpenID-Connect (Keycloak)

Mit der rollenbasierter Domänenabsicherung können einzelne Domänen für authentifizierte Benutzer, basierend auf den ihnen zugeordneten Rollen, ein- bzw. ausgeblendet werden. So werden über spezielle Rollen die Domänen beschrieben, auf die der Zugriff erlaubt sein soll. Alle anderen Domänen werden "ausgeblendet" bzw. sind nicht zugänglich.

Als Paradigma wird dabei die transparente "Perspektive" (oder "View") verwendet: Anfragen zur Domänen-Auflistung werden nur mit den Domänen beantwortet, zu denen es eine Autorisierung gibt. Zugriffsversuche auf andere Domänen werden so beantwortet, als gäbe es diese nicht. So ist es einem Nutzer auch nicht möglich, durch gezielte Anfragen herauszufinden, welche weiteren Domänen in der Instanz vorhanden sind.

Die "Filterung" der Domänen erfolgt im Backend, so dass die Zugriffe über SOAP und das WEB-Frontend entsprechend eingeschränkt werden, sofern diese authentifiziert und mit aktivierter rollenbasierter Domänenabsicherung erfolgen.

Das zweistufige Rollensystem mit Admin- und User-Rollen (vgl. **Abschnitt 11.1**) bleibt von rollenbasierter Domänenabsicherung unberührt und ist komplementär dazu.

 **Hinweis:** Weitere inhaltliche Erläuterungen zur Verwendung und Konfiguration der Domänenspezifischen Rollen und Rechte sind separat unter <https://www.ths-greifswald.de/ttp-tools/domain-auth> dokumentiert.

12 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)

13 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
```

13.1 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 Übereinstimmung 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 13-1** 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 13-1: Exemplarische Anfrage zur Registrierung einer Person.

Die Antwort auf die eben gezeigte Anfrage ist in **Abbildung 13-2** 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 13-2: 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 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.

13.2 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 13-3** 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 13-3: Exemplarische Anfrage zum Suchen einer Person mittels dazugehörigen MPI.

In **Abbildung 13-4** 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 13-4: Antwort auf die Anfrage zum Suchen einer Person mittels des MPIs.

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

13.3 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 13-5** 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 13-5: Anfrage um alle Personendaten einer Domain abzurufen.

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

14 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.

14.1 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 14.1.1**) 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 14.1.2**). Je nach Ähnlichkeit können die verbleibenden Datensätze als Dublette, keine Dublette oder mögliche Dublette klassifiziert werden. In **Abbildung 14-1** ist der Prozess der Registrierung stark vereinfacht dargestellt.

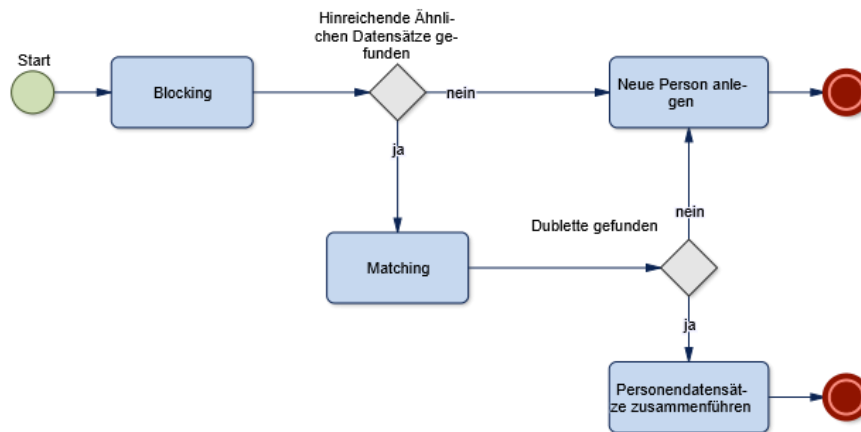


Abbildung 14-1: Vereinfachter Ablauf des Matching-Prozesses.

14.1.1 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 14.4.12**). Das Blocking dient der Steigerung der Performance und verringert insbesondere bei großen Datenbeständen die Dauer eines Abgleichs.

14.1.2 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 14.4.12**) werden können. Mithilfe dieses genaueren Abgleichs kann klassifiziert werden, ob ein Duplikat vorliegt oder nicht¹⁴.

14.2 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.

¹⁴ Tatsächlich findet eine feinere Unterteilung statt (vgl. **Tabelle 7-1**)

Domänen verwalten

Name	Schlüssel	Modus	MPI Identifier-Domäne	Sichere Datenquelle
Demo (aktiv)	Demo	MI	MPI	dummy_safe_source
Demo-Domäne	demo-domain	MI	MPI	Krankenhausinformationssystem

Datenquellen verwalten

Name	Schlüssel
Krankenhausinformationssystem	KIS
dummy_safe_source	dummy_safe_source

Identifizier-Domänen verwalten

Name	Schlüssel	OID
MPI	MPI	1.2.276.0.76.3.1.132.1.1.1

Abbildung 14-2: Das Anzeigen und Editieren der aktuellen Konfiguration einer E-PIX-Domäne ist direkt über das Web-Frontend möglich.

In **Abbildung 14-3** 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. Die angegebene Reihenfolge der Elemente ist dabei einzuhalten. Eine Erläuterung aller Elemente mit Beispielen und validen Wertebereichen folgt im nächsten Abschnitt.



Abbildung 14-3: 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 14-1 sind alle vordefinierten Felder aufgelistet. Hierbei ist zu beachten, dass die Felder der Kontaktdaten nicht für das Matching verwendet werden können.

Tabelle 14-1: Alle im E-PIX definierten Felder.

Feldname	Beschreibung	Beispiel
firstName	Vorname	Anna
middleName	Weitere Vornamen	Lea
lastName	Nachname	Schmidt
birthDate	Geburtsdatum Format: JJJ-MM-TT	1980-03-12
gender	Geschlecht (wird intern auf mittels eines f Buchstaben angegeben) <i>m</i> für male (männlich), <i>f</i> für female (weiblich), <i>o</i> für other (sonstige), <i>u</i> für Unknown (unbekannt) und <i>x</i> für divers	
externalDate	Freies Feld für ein Datum, welches im E-PIX nur gespeichert, aber nicht weiter prozessiert wird Format: JJJ-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	Todesdatum

	Abschnitt 14.4.8). Die Felder haben in der Datenbank unterschiedliche Längen und sind auf diese limitiert:	
	value1-5	max. 50 Zeichen
	value6 und 7	max. 255 Zeichen
	value8 und 9	max. 1.000 Zeichen
	value10	max. 15.000 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	<i>beliebig</i>
vitalStatus	Vitalstatus Unterstützte Werte sind: <i>ALIVE</i> (lebendig), <i>DEAD</i> (verstorben), <i>UNKNOWN</i> (unbekannt)	<i>ALIVE</i>
dateOfDeath	Sterbedatum	2015-03-20

14.3 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 14.4.11**) aufbereitet. Für das Blocking werden die Felder `firstName` und `birthDate` verwendet. Für das Feld `firstName` werden zudem Multiple-Values (siehe **Abschnitt 14.4.12.7**) genutzt. Ein Matching findet mithilfe aller vier Felder statt. Für einen Abgleich wird immer die Levenshtein-Distanz verwendet¹⁵. In **Tabelle 14-2** sind die Felder zur Übersicht dargestellt.

Tabelle 14-2: Verwendete Felder mit Schwellwerten und Wichtung in der Standard-Domänenkonfiguration.

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

14.4 Struktur und Inhalt der Konfiguration

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

14.4.1 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 14-3** sind die zwei Modi im Detail erläutert.

Tabelle 14-3: Unterstützte Matching-Modes

Wert	Beschreibung
<code>MATCHING_IDENTITIES</code>	Bei der Registrierung von Personen wird ein Record Linkage durchgeführt (Verwendung von <code>addPerson</code> nicht möglich). Die

¹⁵ Weitere Vergleichsmöglichkeiten sind implementiert (vgl. **Tabelle 14-10**)

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 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.

14.4.2 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.

14.4.3 mpi-prefix

Die ersten Ziffern im MPI können mithilfe eines Präfixes festgelegt werden. Jeder MPI enthält damit die angegebene Ziffernfolge¹⁶. 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.

14.4.4 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,

¹⁶ Ob das Präfix verwendet wird, hängt davon ab, ob der genutzte MPI-Generator das Präfix berücksichtigt. Der mitgelieferte Generator (`EAN13Generator`) berücksichtigt das Präfix.

wird in jedem Fall, auch wenn in der Domänenkonfiguration anders definiert, eine Notification versendet. In **Tabelle 14-4** sind alle derzeit unterstützten Notifications aufgelistet.

Tabelle 14-4: Unterstützte Notifications im E-PIX.

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 zusammengeführt.

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

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

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

14.4.5 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 14.4.12.6**) 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.

14.4.6 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 14.4.10** zu finden. In **Listing 6** wird exemplarisch die Festlegung des Persist-Modus dargestellt.

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

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

14.4.7 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 14-1 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.

14.4.8 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 14-4**) 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 14-4: Weboberfläche zur Registrierung einer Person. Rechts sind die gemappten Felder dargestellt.

14.4.9 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.

14.4.10 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.

14.4.10.1 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 14-1) und 2) der Bloomfilter aus normalisierten bzw. aus aufbereiteten Werten generiert wird (siehe **Abschnitt 14.4.11**). Der Bloomfilter kann wie andere Felder auch zum Matching verwendet werden. Hierzu stehen entsprechende Vergleichsverfahren zur Verfügung. Im **Abschnitt 14.4.12.6.6** 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 Härtingsverfahren, 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 14-5: Elemente der Bloomfilter-Konfiguration.

Element-Name	Beschreibung	Beispiel
<code>algorithm</code>	Angabe des Algorithmus, welcher das Verfahren zur Erzeugung des Bloomfilters implementiert. Eine Auflistung von den unterstützten Algorithmen ist in Tabelle 14-6 zu finden.	<code>org.emau.icmvc.ttp. deduplication.impl. bloomfilter. RandomHashingStrategy</code> ↔ ↔ ↔
<code>field</code>	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 14-1) empfohlen (je nach Konfiguration).	<code>value8</code>
<code>length</code>	Länge des Bloomfilters in Bits.	1000

<i>ngrams</i>	Länge der N-Gramme, die für die Erzeugung des Bloomfilters verwendet werden. Klassischerweise wird hier ein Wert von 2 angegeben, um Bi-Gramme zu erzeugen.	2
<i>bits-per-ngram</i>	Anzahl der Bits, die pro N-Gramm im Bloomfilter gesetzt werden. Beim Double-Hashing wird von Iterationen gesprochen. Beim Random-Hashing handelt es sich um die Anzahl der generierten Zufallspositionen.	25
<i>fold</i>	Der E-PIX unterstützt ein XOR-Folding von Bloomfiltern nach Schnell et al. ¹⁷ . 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, ...
<i>alphabet</i>	Das Alphabet, welches beim Random-Hashing berücksichtigt werden soll (nur erforderlich, wenn das Random-Hashing verwendet wird).	ABCDEF12345-
<i>balanced</i>	Der E-PIX unterstützt das Generieren von Balanced-Bloomfiltern (Schnell et al. ¹⁸). Das Element <i>balanced</i> enthält ein Feld <i>seed</i> , welches einen Zahlenwert enthält. Dieser stellt den Seed-Wert des Zufallsgenerators dar. Wird dieses Element (<i>balanced</i>) nicht angegeben, wird kein Balanced-Bloomfilter erzeugt. Der Balanced-Bloomfilter führt zu einer Verdopplung der resultierenden Bloomfilter-Länge.	462945623209
<i>source-field</i>	Jeder Bloomfilter kann aus einem oder mehreren Feldern zusammengesetzt werden. Dabei wird je Feld (Element: <i>field</i> (enthält Feldnamen, siehe Tabelle 14-1)) der Wert entsprechend gehashed. Beim Random-Hashing kann pro Feld ein Seed-Wert (Element: <i>seed</i> (enthält einen Zahlenwert)) gesetzt werden. Beim Double-Hashing kann ein Salt auf Basis einer statischen Zeichenkette (Element:	

¹⁷ https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3527984

¹⁸ <https://ieeexplore.ieee.org/document/7836669>

`salt-value` (enthält eine feste Zeichenkette (z.B.: `a3ghd5o36#sz3`)) oder dynamisch auf Basis eines anderen Feldes (Element: `salt-field` (enthält Feldnamen, siehe

Tabelle 14-1)) der Identität gesetzt werden.

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

Tabelle 14-6: Unterstützte Algorithmen zur Generierung von Bloomfiltern.

Algorithmus	Beschreibung
<code>org.emau.icmvc.ttp.deduplication.impl.bloomfilter.RandomHashingStrategy</code>	Random Hashing ¹⁹
<code>org.emau.icmvc.ttp.deduplication.impl.bloomfilter.DoubleHashingStrategy</code>	Double Hashing ²⁰
<code>org.emau.icmvc.ttp.deduplication.impl.bloomfilter.DoubleHashingStrategyFaster</code>	Optimierte Variante vom Double Hashing (Nicht Kompatibel mit <code>DoubleHashingStrategy</code>)

¹⁹ 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.

²⁰ 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.

14.4.11 preprocessing-config

Mithilfe des pre-processing 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 14.4.11.2** wird das Element `field-name`, in **Abschnitt 14.4.11.3** wird das Element `simple-transformation-type` und in **Abschnitt 14.4.11.4** das Element `complex-transformation-type` erläutert.

14.4.11.1 *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` verarbeitet²¹.

14.4.11.2 *field-name*

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

Tabelle 14-1 angegeben.

14.4.11.3 *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

²¹ Die Festlegung der Reihenfolge wird demnächst implementiert.

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.

14.4.11.4 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 14-7** 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 14-7: Unterstützte Transformationen für `complex-transformation-type`.

Transformation	Beschreibung	Beispiel
<code>ToUpperCaseTransformation</code>	Alle Kleinbuchstaben werden durch Großbuchstaben ersetzt.	Anna → ANNA
<code>CharsMutationTransformation</code>	Ersetzt Umlaute.	München → Muenchen
<code>TrimTransformation</code>	Entfernt führende und nachfolgende Leerzeichen.	_ _An_ _na_ → An_ _na
<code>CharNormalizationTransformation</code>	↩ 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.emaui.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.

14.4.11.5 simple-filter-type

Mit `simple-filter-type` kann ein Alphabet (`pass-alphabet`) bestimmt werden. Alle Zeichen die davon abweichen, werden durch das angegebene Ersatz-Zeichen (`replace-character`) ersetzt. Ist das Ersatz-Zeichen leer, so werden die Zeichen entfernt, die nicht im Alphabet enthalten sind. In **Listing 14** ist ein einfaches Beispiel zum Entfernen ungültiger Zeichen dargestellt.

```

...
<simple-filter-type xsi:type="ma:SimpleFilter">
  <pass-alphabet>ABCDEFGHIJKLMNOPQRSTUVWXYZ</pass-alphabet>
  <replace-character></replace-character>
</simple-filter-type>
...

```

Listing 14: XML-Code zur Definition eines Filters, zum Entfernen aller Zeichen, die nicht Teil des Alphabets A-Z sind.

14.4.12 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 14-8** sind die empfohlenen und Standard-Schwellwerte dargestellt.

Tabelle 14-8: Empfohlene und Standard-Schwellwerte für *Automatic Match* und *Possible Match*.

Schwellwert	Wert (gemäß Standard-Konfiguration siehe Abschnitt 14.3)	Standardwert (wenn nicht angegeben)
<code>threshold-automatic-match</code>	14,5	20
<code>threshold-possible-match</code>	2,99	4

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

14.4.12.1 *threshold-possible-match*

Mit dem Element `threshold-possible-match` kann der Schwellwert für *Possible Matches* (vgl. **Tabelle 14-8**) definiert werden. Überschreitet die ermittelte Wahrscheinlichkeit den angegebenen Wert (und unterschreitet den Schwellwert `threshold-automatic-match`), so wird der Match-Typ *Possible Match* als Ergebnis des Record Linkages zurückgegeben. In **Listing 151514** ist die Definition des Schwellwert dargestellt.

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

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

14.4.12.2 *threshold-automatic-match*

Mit dem Element `threshold-automatic-match` kann der Schwellwert für *Automatic Matches* (vgl. **Listing 161615**) 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 161615** ist die Definition des Schwellwert dargestellt.

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

Listing 16: XML-Code zur Definition des Schwellwerts für *Automatic Matches*.

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

Eine automatische Zusammenführung kann auf *Perfect Matches* beschränkt werden. Fälle mit sehr hoher Übereinstimmung, 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.

14.4.12.3 *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 14-9** entnommen werden.

Tabelle 14-9: 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	<code>true</code>	Ja	Fehler: Ein Identifier darf nur einer Person pro Domäne zugeordnet sein.
2	<code>false</code>	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 171716** ist exemplarisch die Definition dargestellt.

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

Listing 17: XML-Code zur Definition des *use-cemfim*-Wertes.

14.4.12.4 *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 181817** ist exemplarisch die Definition dargestellt.

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

Listing 18: XML-Code zur Definition der Anzahl registrierter Personen, ab denen der E-PIX Multithreading verwendet.

14.4.12.5 *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 191918** ist die exemplarische Definition der Anzahl der verwendeten Threads dargestellt.

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

Listing 19: XML-Code zur Definition der Anzahl der verwendeten Threads.

14.4.12.6 *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. Dabei ist zu beachten, dass wenn nur ein Feld zu Matching genutzt wird, dass das Gewicht auf 100 gesetzt wird. Werden mehrere Felder verwendet, werden die Felder im Verhältnis ihres Gewichts in die Berechnung einbezogen. In **Listing 202019** 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 20: XML-Code zur exemplarischen Konfiguration eines Felders, welches zum Matching verwendet wird.

14.4.12.6.1 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 14-1 angegeben. In **Listing 212120** ist exemplarisch der Wert „gender“ angegeben, wenn das Geschlecht z.B. für das Blocking verwendet werden soll.

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

Listing 21: XML-Code zur Definition des Feldes für das Record Linkage.

14.4.12.6.2 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 222221** wird exemplarisch ein Schwellwert definiert.

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

Listing 22: XML-Code zur Definition eines Schwellwertes für das Blocking von einem Feld.

14.4.12.6.3 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 232322** ist die Definition von `blocking-mode` exemplarisch für Zahlenvergleiche dargestellt.

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

Listing 23: XML-Code zur Definition der Blocking-Vergleichsmethode.

14.4.12.6.4 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 242423** ist exemplarisch eine Schwelle definiert.

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

Listing 24: XML-Code zur Definition eines Schwellwertes für das Matching von einem Feld.

14.4.12.6.5 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 252524** ist exemplarisch eine Wichtung angegeben.

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

Listing 25: XML-Code zur Wichtung eines Feldes.

14.4.12.6.6 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 14-10** 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 14-10: Unterstützte Algorithmen für das Matching.

Algorithmus	Beschreibung
<code>ColognePhoneticAlgorithm</code>	Vergleicht zwei Werte nach ihrem Sprachklang. Die Nachnamen Maier, Meyer und Meier würden beispielsweise als gleich gewertet werden.
<code>DeterministicAlgorithm</code>	Vergleicht zwei Werte auf exakte Gleichheit. Bei exakter Gleichheit zweier Werte ist das Ergebnis 1, bei einer Abweichung 0.
<code>LevenshteinAlgorithm</code>	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.
<code>SorensenDiceCoefficientCoded</code>	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 262625** wird exemplarisch die Definition eines Algorithmus zum Abgleich von einem Feld definiert.

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

Listing 26: XML-Code zur Definition des Algorithmus für das Matching.

14.4.12.7 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. Hierzu kann ein Separator definiert werden, anhand dessen die Teil-Zeichenketten ermittelt werden. In **Listing 272726** 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 27: XML-Code zur Definition eines `multi-value`-Feldes.

14.4.12.7.1 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 282827** ist die Definition eines Leerzeichens als Separator dargestellt.

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

Listing 28: XML-Code zur exemplarischen Definition eines Leerzeichens als Separator eines `multi-value`-Feldes.

14.4.12.7.2 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 303029** ist exemplarisch gezeigt, wie ein Wert hierfür definiert wird.

14.4.12.7.3 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 29: XML-Code zur exemplarischen Definition des `penalty-one-short`-Wertes.

14.4.12.7.4 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 303029** ist exemplarisch gezeigt, wie ein Wert hierfür definiert wird.

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

Listing 30: XML-Code zur exemplarischen Definition des `penalty-both-short`-Wertes.

15 Optimierungen

15.1 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 erhöht 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 Applikationsservers 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 Deployment-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>
```


15.2 Optimierungen bei Betrieb ohne Docker

Wird entgegen der hier beschriebenen 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.

15.2.1 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.

15.2.2 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.

15.2.3 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 anzupassen. 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>
```

16 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.

17 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>