PHI und Pseudonymisierung in GeMTeX

Stefan Schulz, Averbis GmbH

Luise Modersohn, TU München

8.8.2023

ÜBERBLICK

- Konsistente Pseudonymisierung
  - Konzept
    - Unterschied zum üblichen Begriff der Pseudonymisierung von Datensätzen
    - Unterschied zur Anonymisierung von Texten
  - Ziele
  - Umsetzung
- Einbettung der konsistenten Pseudonymisierung in den GeMTeX-Workflow

KONSISTENTE PSEUDONYMISIERUNG KLINISCHER DOKUMENTE

averbis
text analytics

- "Pseudonymisierung“ heißt meistens Ersatz der Identifikatoren eines Datensatzes mit Möglichkeit der Re-Identifizierung seitens Vertrauensstelle
- „Anonymisierung“: aus-X-en von PHIs
- In GemTeX keine Anonymisierung, dafür Ersatz aller PHI-getaggten Strings durch geeignete Ersatz-Strings
- Ziele:
  - Sicherheit: keine Re-identifizierung möglich
  - Plausibilität: das Ergebnis soll von einem Originaldokument in der Struktur nicht unterscheidbar sein
  - Konsistenz: Durch Pseudonymisierung dürfen innerhalb von Dokumenten eines Patienten keine Zusammenhänge (Personen, Zeitangaben) verloren gehen

HIPPAA PHI (Protected Health Information)

Names (Full or last name and initial)
All geographical identifiers smaller than a state (…)
Phone Numbers
Fax numbers
Email addresses
Social Security numbers
Medical record numbers
Health insurance beneficiary numbers
Account numbers
Certificate/license numbers
Vehicle identifiers
Device identifiers and serial numbers;
Web Uniform Resource Locators (URLs)
Internet Protocol (IP) address numbers
Biometric identifiers, including finger, retinal and voice prints
Full face photographic images and any comparable images
Any other unique identifying number, characteristic, or code (…)

Original

Krankenhaus der Samariter Holzhausen

Röntgenabteilung, Vorstand Prim. Univ. Prof. Dr.Dr. Gotthard Vogler

CT Abdomen und kl. Becken

Name: Mustafa Üstün, * 21.06.67

Aufnahmezahl: 1933309807

Abteilung: Chirurgie

Station: A31. OG. Viszeralchirurgie B /

Zi: 119

dikt. Arzt: OA Dr. Huber Karina

WinA. 06/07/2011

Getaggt mit DE-ID der Averbis Health Discovery

<location>Krankenhaus der Samariter</location>
<location>Holzhausen</location>

Röntgenabteilung, Vorstand <name>Prim. Univ. Prof. Dr.Dr. Gotthard Vogler</name>

CT Abdomen und kl. Becken

Name: <name>Mustafa Üstün</name>, * <date>21.06.67</date>

Aufnahmezahl: <id>1933309807</id>

Abteilung: Chirurgie

Station: <division>A31. OG. Viszeralchirurgie B</division> /

Zi: 119

dikt. Arzt: <name>OA Dr. Huber Karina</name>

WinA. <date>06/01/2011</date>

Getaggt

<location>Krankenhaus der Samariter</location>
<location>Holzhausen</location>

Röntgenabteilung, Vorstand <name>Prim. Univ. Prof. Dr.Dr. Gotthard Vogler</name>

CT Abdomen und kl. Becken

Name: <name>Mustafa Üstün</name>, * <date>21.06.67</date>

Aufnahmezahl: <id>1933309807</id>

Abteilung: Chirurgie

Station: <division>A31. OG. Viszeralchirurgie B</division> /

Zi: 119

dikt. Arzt: <name>OA Dr. Huber Karina</name>

WinA. <date>06/01/2011</date>

Anonymisiert

XXXXXXXXXX XXXXXXXXXX

Röntgenabteilung, Vorstand Prim. Univ. Prof. Dr.Dr. XXXXXXXX XXXXXXXXXXXXXX

CT Abdomen und kl. Becken

Name: XXXXXXX XXXXXXXXXX, * X.X.X

Aufnahmezahl: XXXXXXXXXXXX

Abteilung: Chirurgie

Station: XXXXXXXXX XXXXXXXXXXX

Zi: 119

dikt. Arzt: OA Dr. XXXXXXXXX XXXXXXXXX

WinA. XX/XX/XXXX

Getaggt

<location>Krankenhaus der Samariter</location>
<location>Holzhausen</location>

Röntgenabteilung, Vorstand <name>Prim. Univ. Prof. Dr.Dr. Gotthard Vogler</name>

CT Abdomen und kl. Becken

Name: <name>Mustafa Üstün</name>, * <date>21.06.67</date>

Aufnahmezahl: <id>1933309807</id>

Abteilung: Chirurgie

Station: <division>A31. OG. Viszeralchirurgie B</division> /

Zi: 119

dikt. Arzt: <name>OA Dr. Huber Karina</name>

WinA. <date>06/01/2011</date>

Pseudonymisiert als separater Prozess

Kantonsspital Friedrichshafen

Röntgenabteilung, Vorstand Prim. Univ. Prof. Dr.Dr. Gerhard Voigtländer

CT Abdomen und kl. Becken

Name: Manuel Überreuter, * 1.07.69

Aufnahmezahl: 9983209971

Abteilung: Chirurgie

Station: Station Sauerbruch

Zi: 119

dikt. Arzt: OA Dr. Heilmann Kristina

WinA. 16/07/2013

VERSCHIEBEN VON ZEITANGABEN UNTERSCHIEDLICHER GRANULARITÄT

averbis
text analytics

Einheit	Algorithmus	Beispiel (offset = 300)
Tag	$d_{pseudo} = d_{orig} + offset$	11.03.2021 –> 05.01.2022
Monat	$d_{median} = floor\ (\ median\ (\ days\ (\ m_{orig}\ )))$ $m_{pseudo} = month\ (\ d_{orig} + offset\ )$	3/2021 –> 16.03.2021 16.03.2021 –> 10.01.2022 –> 1/2022
Jahr	$d_{median} = floor\ (\ median(days(y_{orig}\ )))$ $y_{pseudo} = year\ (\ d_{orig} + offset\ )$	2021 –> 01.07.2021 01.07.2021 –> 27.04.2022 –> 2022
Nominale Zeitangaben		
Feiertage	Ersatz durch unspezifische Angaben oder Weglassen	"am Karfreitag, den 4. April operiert" -> "am Karfreitag, den 3. Januar" operiert -> "am 3. Januar operiert" "nach Pfingsten" -> "nach dem Feiertag"
Quartale, Jahreszeit	Analoges Vorgehen zu Monaten	Sommer 2021 –> 07.08.2021 07.08.2021 –> 03.06.2022 –> Frühjahr 2022

PERSONENNAMEN: NORMALISIERUNG / KATEGORISIERUNG

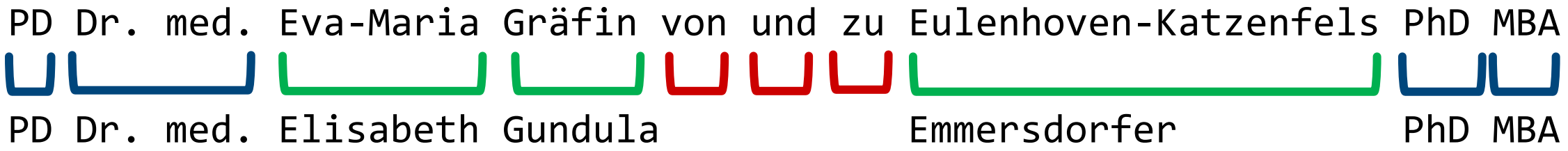
averbis
text analytics

Namensbestandteile		Beispiel
Trennzeichen	Leerzeichen	`Anna Osler-> ('Anna', 'Osler')`
Bindestrich	Kein Trennzeichen	`Eva-Maria Rau ->('Eva-Maria','Rau')`
1. Buchstabe klein	Ignoriert	`'van Beethoven' -> 'Beethoven'`
Token bis *ceil*(*median*(*n*))	Vornamen (V), dann Nachnamen (N)	`'Kim Yong Il Park Un'` `V: 'Kim Yon Il', N: 'Park Un'`
Namenszusätze		
Mit Punkt	Ignoriert	`'Dr.', 'Prof.', 'Dipl.-Ing','Jr.'`
Ohne Punkt	Ignoriert (aus Lexikon)	`'OA', 'PD', 'PhD', 'MBA'`
Repetitionen	Ignoriert	`'Dr. Dr. Dr.' -> 'Dr'`

NAMEN: PSEUDONYMISIERUNG

averbis
text analytics

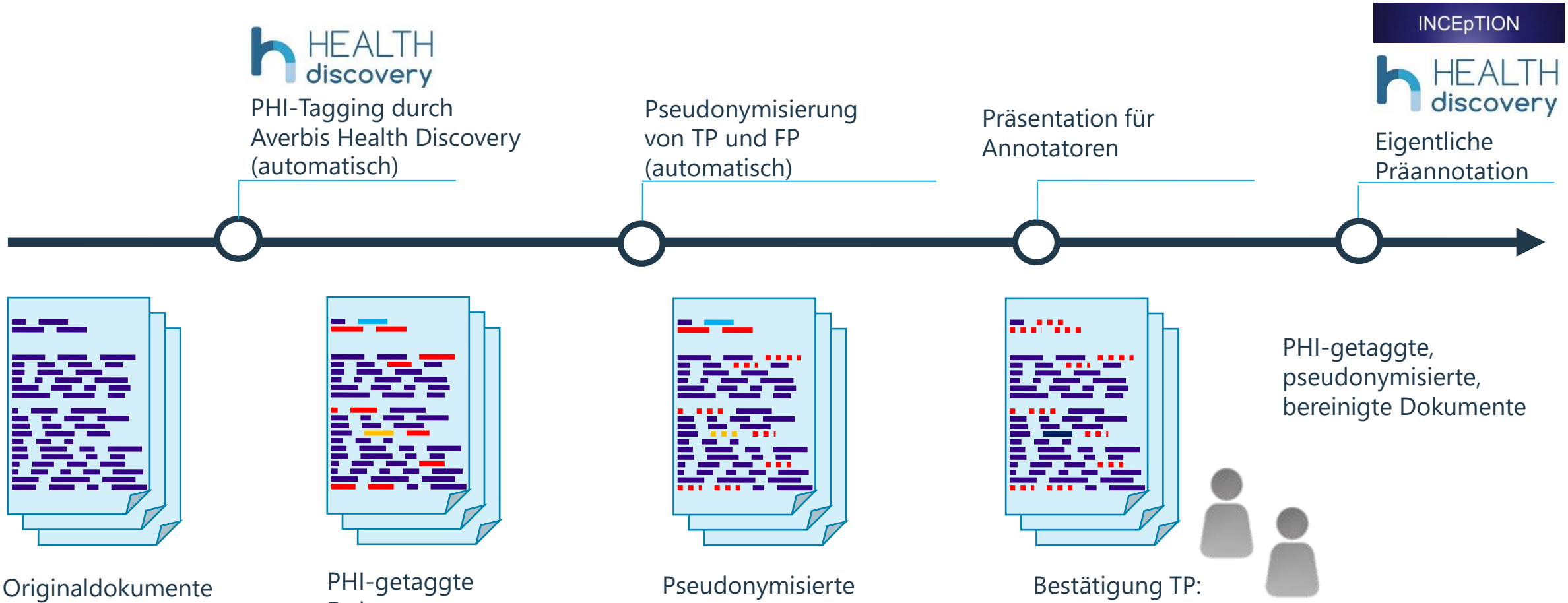
Namen		Beispiel
Vornamen	Vornamenlexikon, indexiert nach Initiale und Geschlecht	Zufallszahl r W = (…,'Andrea', 'Agnes', …) 'Anna' -> w_r , z.B. 'Anna' -> 'Agnes'
Nachnamen	Nachnamenlexikon, indexiert nach Initiale	Zufallszahl r N = (…,'Emmerich', 'Emmersdorfer', 'Eils', …) 'Eberhard' -> n_r , z.B. 'Eberhard' -> 'Emmerich'
Kombina-tionen	Ausnahmen für seltene Initialien-Kombinationen	'Özgan Ötztürk' -> ‚Florian Österreicher'
Institutionen	Lexikon Pseudo-Lokalisationen Abgleich mit Terminologie	'LKH Salzburg' -> 'Scheuermann-Klinik' 'in HNO verlegt' -> 'in HNO verlegt'



KONSISTENTE PSEUDONYMISIERUNG IN GEMTEX

- Warum Pseudonymisierung und nicht Anonymisierung (aus-X-en)
  - PHIs (insbesondere Namen, Vornamen, Institutsbezeichnungen, Datumsangaben) sind prägende Bestandteile von klinischen Dokumenten
  - Durch Anonymisierung
    - Werden Dokumente in ihrer Struktur wesentlich verfälscht
    - Gehen zeitliche Bezüge verloren
    - Gehen Bezüge zwischen Dokumenten zu demselben Patienten verloren
    - Werden falsch negative PHIs leichter erkennbar (echte Patientennamen zwischen ausge-x-ten Inhalten fallen stärker auf als echte Namen zwischen Pseudonamen)
  - Die Generierung von Pseudo-PHIs erfordert allerdings einen eigenen Verarbeitungsschritt, der nicht durch die bisherigen Tools bewerkstelligt wird

VORSCHLAG


HEALTH discovery
PHI-Tagging durch Averbis Health Discovery (automatisch)
Pseudonymisierung von TP und FP (automatisch)
Präsentation für Annotatoren
INCEpTION
HEALTH discovery
Eigentliche Präannotation
PHI-getaggte, pseudonymisierte, bereinigte Dokumente
Originaldokumente (sensibel, mit Originalnamen, Ids und Datumsangaben
PHI-getaggte Dokumente
TP: TN:
FN: FP:
Pseudonymisierte Dokumente
pseudo-TP:
pseudo FP:
Bestätigung TP:
Identifikation: FP: Rückgängigmachen (erfordert Einsicht in Original) oder plausible Korrektur
Identifikation: FN: Nach-Pseudonymisierung (erfordert Einsicht in Regeln)

DISKUSSION

- Einsatz der Health Discovery in zwei Phasen: (i) PHI-Erkennung, (ii) Vorannotation der pseudonymisierten und korrigierten Dokumente
- Notwendigkeit eines von den bisherigen Tools separaten Pseudonymisierungstools
  - Regeln (z.B. Datums-Offset)
  - Ressourcen (Namenslexika)
- Der beschriebene manuelle Verarbeitungsschritt erfordert sowohl Korrektur der Inhalte als auch der Annotationen
  - Ist dies durch Inception unterstützbar?
  - Müssen alle Pseudo-PHI-Strings die Länge der Original-PHI-Strings haben, um Verschiebungen zu vermeiden?