MEDIZIN INFORMATIK INITIATIVE

GeMTeX

Terminologie-Annotationen deutscher Texte im Kontext GeMTeX

Martin Boeker, Markus Löffler, Christina Lohr, Frank Meineke, Luise Modersohn

MEDIZIN INFORMATIK INITIATIVE

2024-09-09

GMDS 2024 Dresden

MRI TUM

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Zielsetzung GeMTeX

- Deutsches medizinisches (klinisches) Referenz-Korpus der MII
- (Prospektive) Textdaten als Ressource für die Forschung
  - Semantische Goldstandard Annotationen
  - Trainierte Sprachmodelle
  - Algorithmische Auswertung
- Nutzung von NLP im Rahmen der Datenintegrationszentren (MII/NUM)
- Initialisierung von Folgevorhaben
  - Demonstration von Vorteilen der semantischen Textanalyse für die Krankenversorgung

Geringe Durchsetzung von NLP an deutschsprachigen klinischen Texten

MEDIZIN INFORMATIK INITIATIVE

GeMTeX

Wichtigste Ursachen

1. Fehlenden ***verfügbare*** deutschsprachigen klinische Korpora
   a. Rechtliche Grundlagen
2. Wenige übersetzte Vokabularien/Terminologien
   a. aber: Interfaceterminologien verfügbar
3. Werkzeuge nicht optimal an deutsche Sprache angepasst

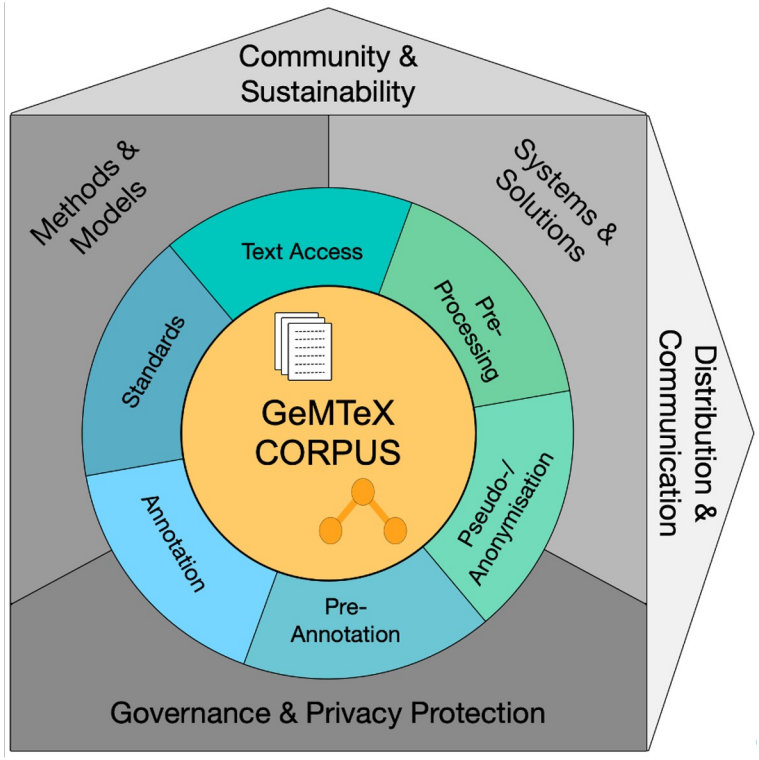
GeMTeX Partner

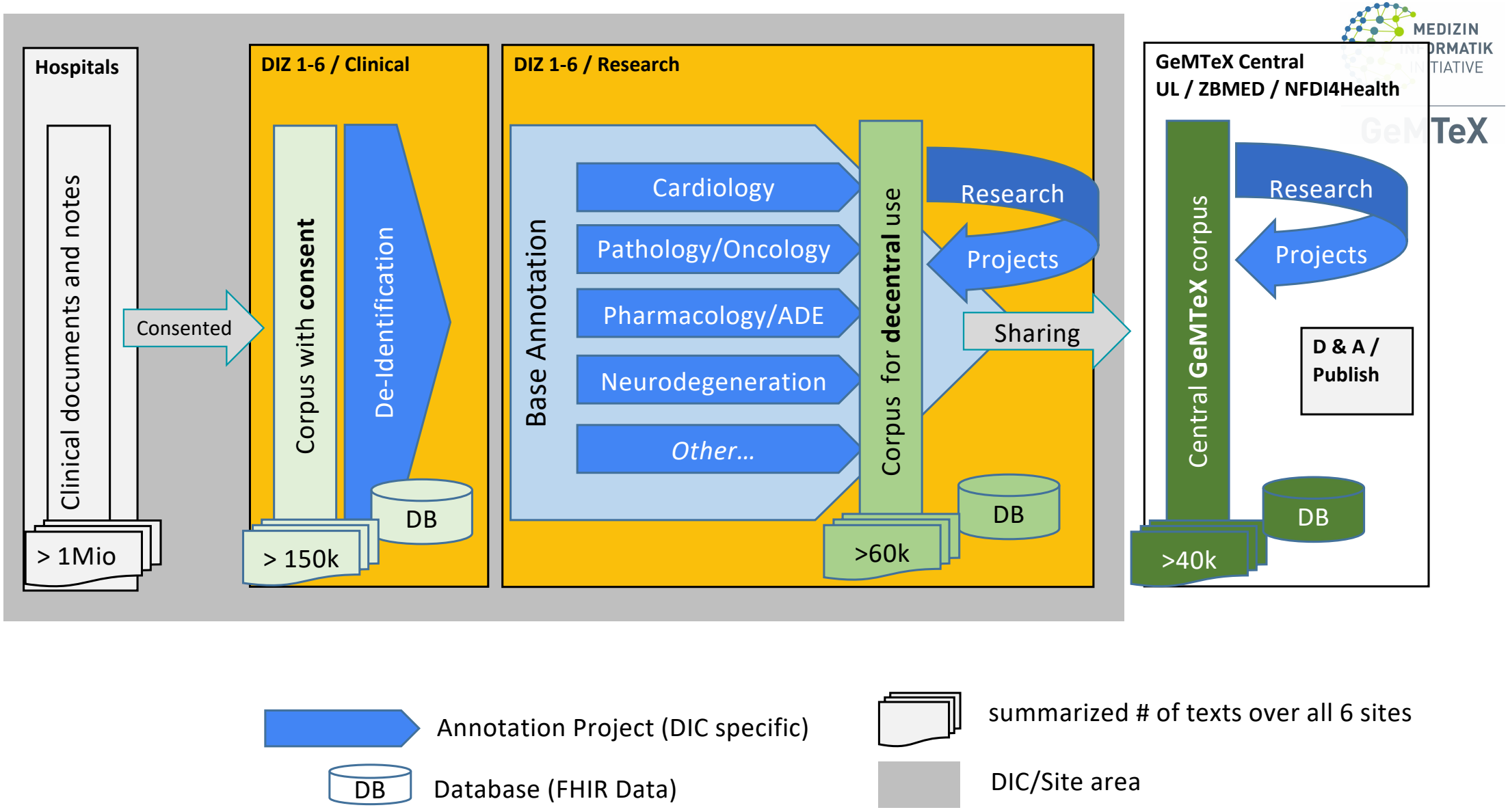
- Technische Universität München
- Universität Leipzig/ Universitätsklinikum Leipzig
- TU Darmstadt
- Universitätsmedizin Essen
- Charité Berlin
- Universitätsklinikum Erlangen
- Universitätsklinikum Dresden
- Universitätsklinikum Heidelberg
- Universität Münster
- Hasso-Plattner-Institut
- Medizinische Hochschule Hannover
- Ludwig-Maximilians-Universität München
- Informationszentrum Lebenswissenschaften ZB MED
- Universitätsklinikum Tübingen
- Averbis GmbH
- ID Berlin
- Medizinische Universität Graz
- Friedrich-Schiller-Universität Jena

- 16 Partner
  - Integration der NWG NLP DE.xt
- 2 assoziierte Partner
- Förderung 6.8 Mio. €
- Laufzeit 3.5 (3) Jahre

Aufbau eines (deutschsprachigen) klinischen Korpus

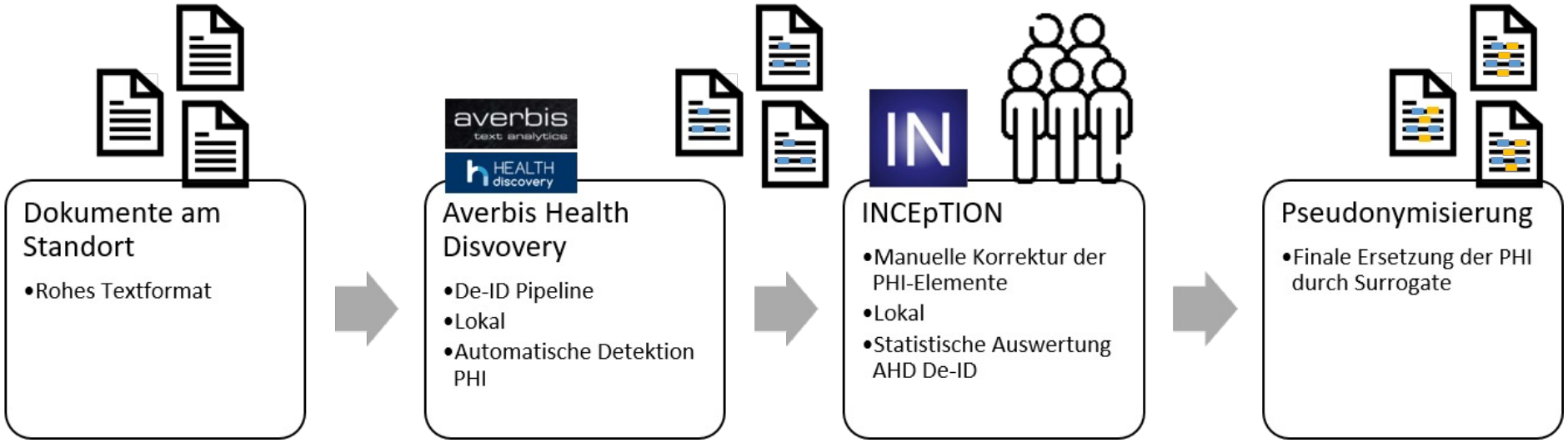
- Ausleitung von Textdokumenten und textuellen Inhalten aus den KIS
  - in allen Formen
  - Tooling wird zur Verfügung gestellt
- De-Identifikation/Anonymisierung
- Annotation
- Nutzungs-Integration
  - lokal
  - zentral aggregiert
  - föderiertes Lernen und Modellintegration
- Governance und Privatheit
- Evaluation des Korpus
- Standardisierung der Prozesse und Repräsentation



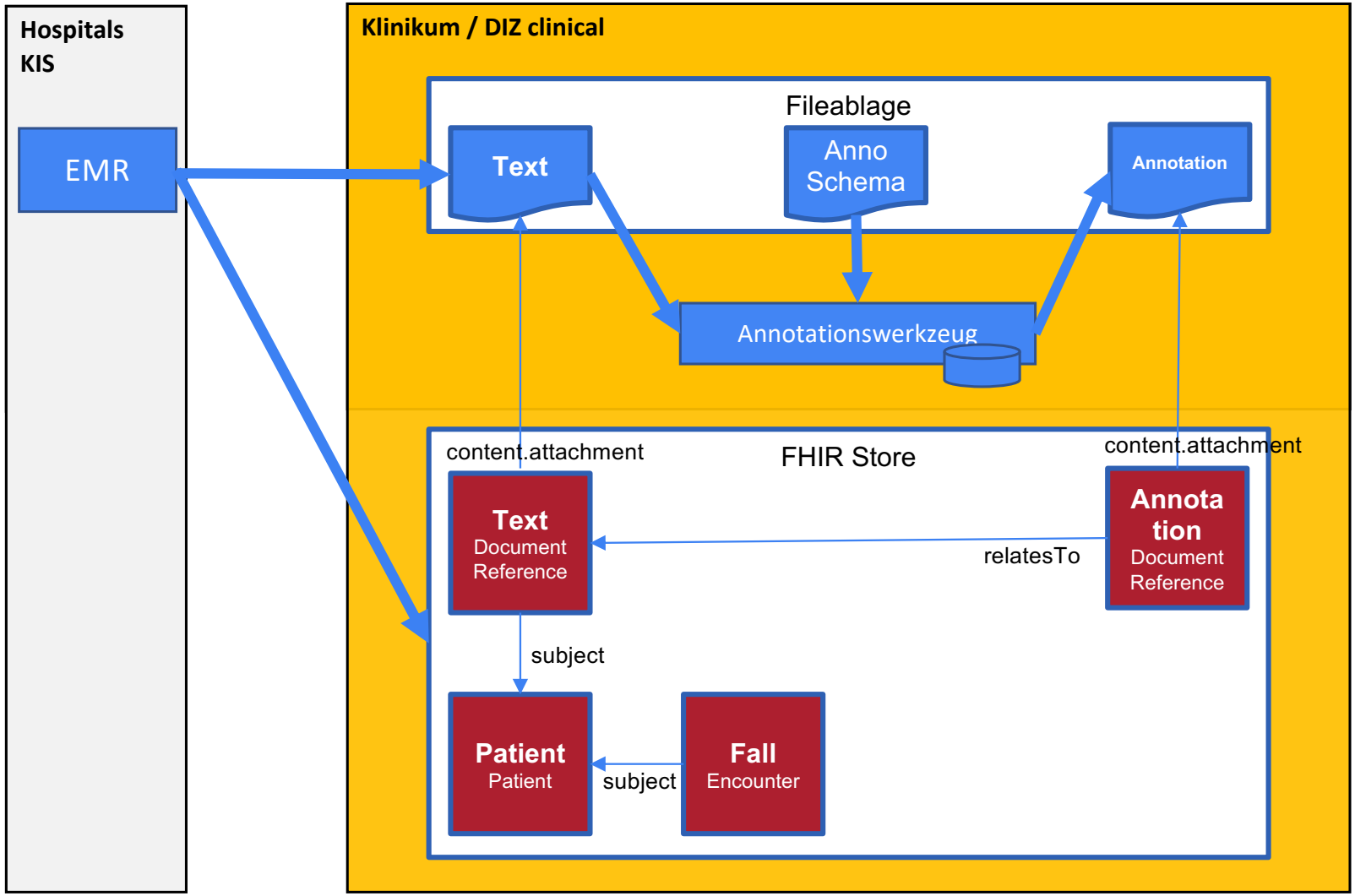

Hospitals
Clinical documents and notes
> 1Mio
Consented
DIZ 1-6 / Clinical
Corpus with consent
De-Identification
DB
> 150k
DIZ 1-6 / Research
Base Annotation
Cardiology
Pathology/Oncology
Pharmacology/ADE
Neurodegeneration
Other…
Corpus for decentral use
DB
>60k
Research
Projects
Sharing
GeMTeX Central
UL / ZBMED / NFDI4Health
Central GeMTeX corpus
Research
Projects
D & A / Publish
DB
>40k
Annotation Project (DIC specific)
DB
Database (FHIR Data)
summarized # of texts over all 6 sites
DIC/Site area

Systemaufbau an Annotations-Standorten

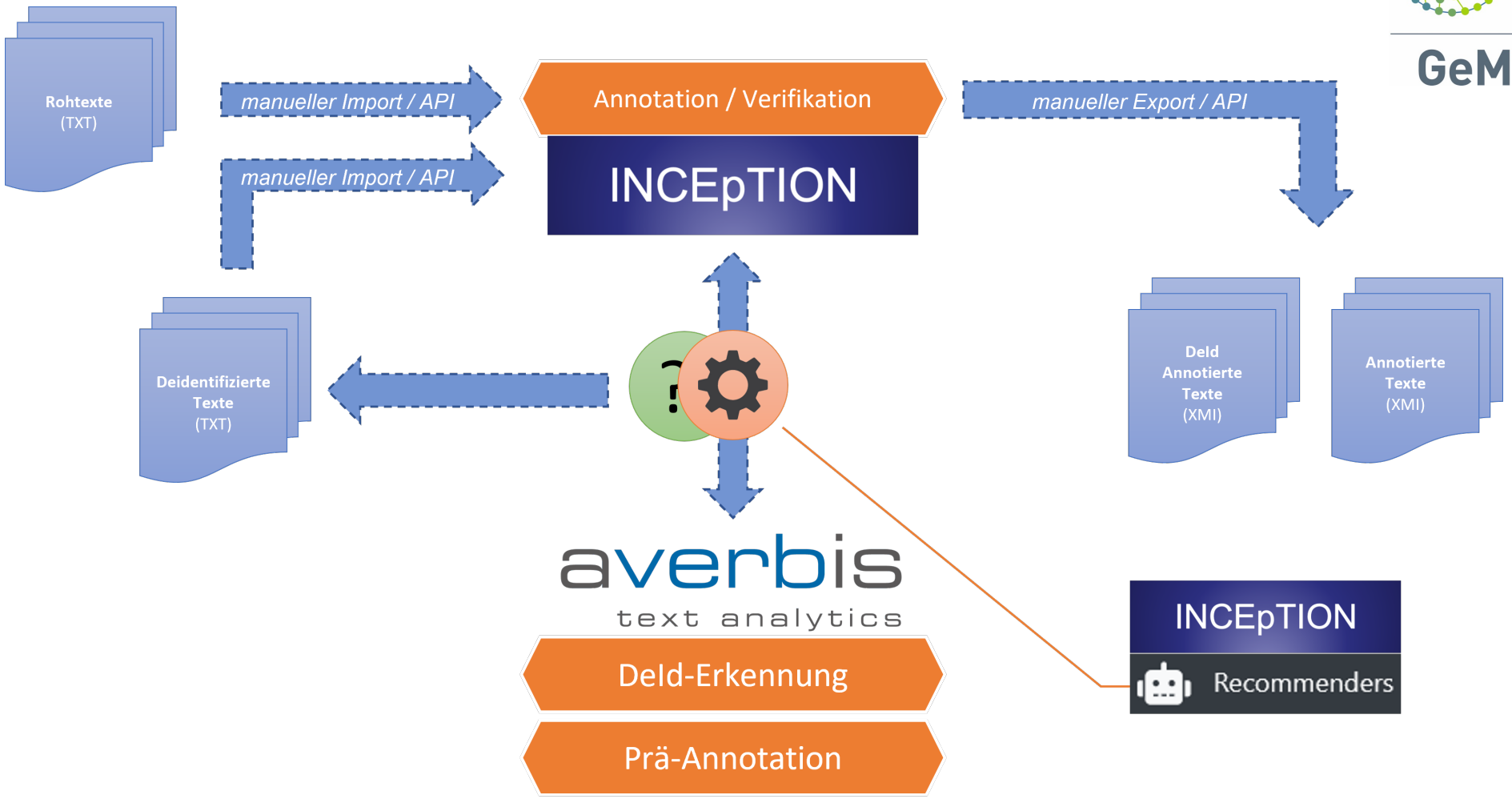
- Bereitstellung Averbis Health Discovery
- Bereitstellung INCEpTION Annotationsplattform
- Bereitstellung sichere Arbeitsumgebungen für Annotatoren
    - Möglichst mit remote Zugriff über VPN



Hinweis: Vortrag von Christina Lohr in der Session NLP und LLM 2024-09-09 um 16:00


Hospitals
KIS
EMR
Klinikum / DIZ clinical
Fileablage
Text
Anno
Schema
Annotation
Annotationswerkzeug
content.attachment
FHIR Store
content.attachment
Text
Document
Reference
relatesTo
Annota
tion
Document
Reference
subject
Patient
Patient
subject
Fall
Encounter

IMISE Referenzplattform



MEDIZIN INFORMATIK INITIATIVE

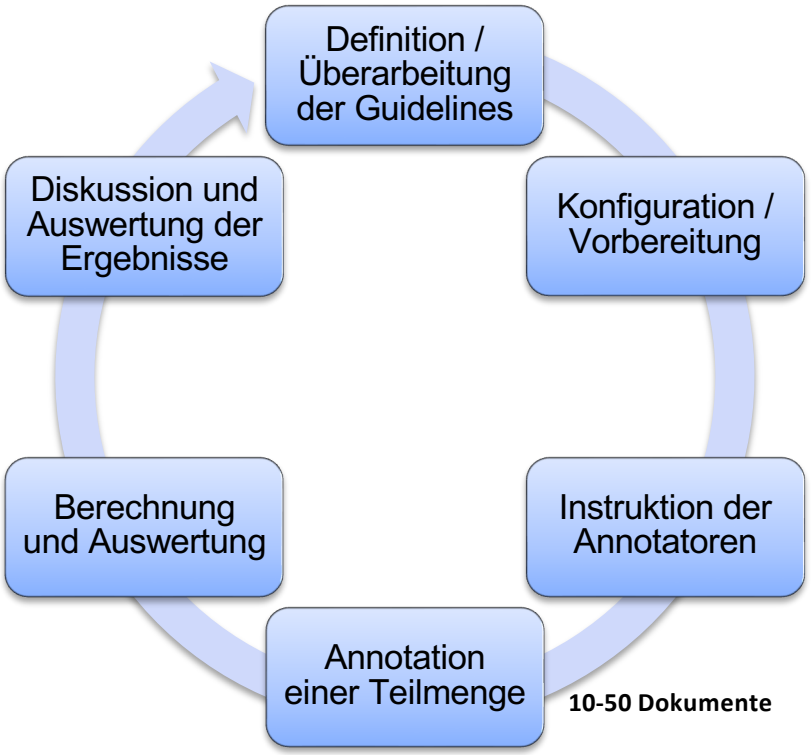
GeMTeX

Annotationsmethodik

- Aufbauend auf bestehenden Methoden/Werkzeugen
  - Annotationsguidelines – basierend auf bestehenden GL (AIDAVA, https://www.aidava.eu/), Vorarbeiten aus Jena (JULIE Lab) & Partnern aus GeMTeX
  - Annotationsterminologie – SNOMED CT
  - Automatisierte Vorannotation – AHD und spez. Projekte (HD, HPI)
  - Annotationseditor (INCEpTION)
- Annotation mit Medizinstudenten an den Standorten
  - Teams von bis zu 10 Studierenden erforderlich
    - Rekrutierung und Nachhaltigkeit der Teams sind
- Annotations-GL (noch nicht publiziert) und Schulungsmaterialien
- Qualitätsprüfung Annotation und De-Identifikation

Schulz, S. et al. (2023). Towards principles of ontology-based annotation of clinical narratives https://ceur-ws.org/Vol-3603/Paper4.pdf
Kolditz T, Lohr C, …, Modersohn L, et al. Annotating German Clinical Documents for De-Identification. MEDINFO 2019: Health and Wellbeing e-Networks for All. 2019;203–7.
Lohr C, Modersohn L, … Hahn U. An Evolutionary Approach to the Annotation of Discharge Summaries. https://ebooks.iospress.nl/doi/10.3233/SHTI200116
Hahn U,…, Lohr C, Löffler, Markus. 3000PA—Towards a National Reference Corpus of German Clinical Language. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. 2018

Annotationszyklus



F1-Score
Krippendorff‘s alpha
Kohen‘s kappa

Mindestens 2 Personen

10-50 Dokumente

Typenebene

- Entitäten, token-of-interest, etc.
- Token span bzw. offset
  (Start und Ende im Text absolut)
- Kleinere Menge von Kategorien

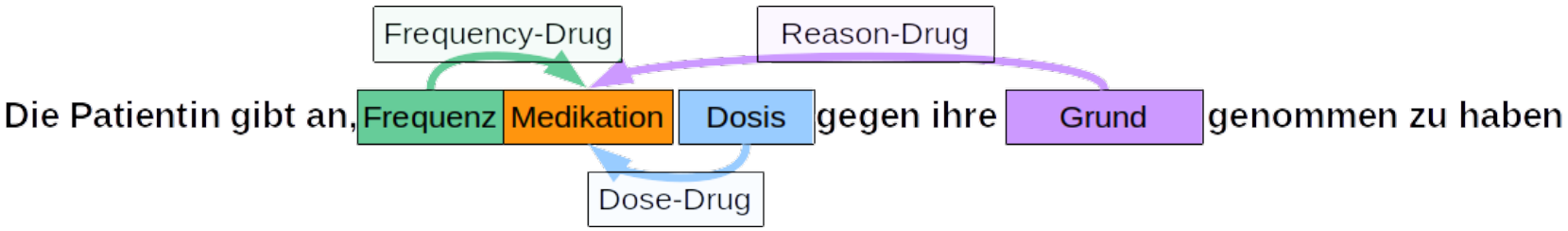
- **Wichtig**: unabhängig voneinander
  - Menge von Dosierungen
  - Menge von Namen, Adressen, etc.

- Beispiele:
  - PHI Kriterien (Name, Vorname, etc.)
  - Zeitangaben (Datum, Zeitspannen, etc.)
    - Vorgestern
    - Zwischen den Jahren
  - Abstraktere Kategorien
    - Typisierung
    - Spezifikationen

Relationsebene

- Relationen und anderweitige Verbindungen
- Unterschiedliche Komplexitätsklassen:

1. „Eindeutige“ Relationen: X[Dosis] *ist Dosierung von* Y[Medikation]
2. „Erwähnte“ Relationen: X[Protein] *methyliert* Y[Gen]
3. „Kontextuelle“ Relationen: X[Datum, Event] *fand statt vor* Y[Datum, Event]



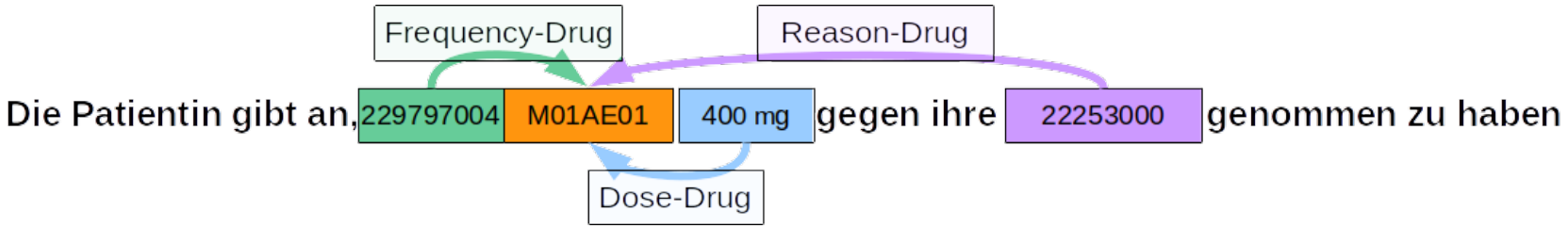
Groundings- und Normalisierungsebene

Grounding

- Abbildung auf Klassifikationssysteme und Terminologien
  - SNOMED CT
  - ICD 10/11

Normalisierung und Vereinheitlichung

- Messwerte in SI-Einheiten
- Datumsangaben in ein eindeutiges Format



GeMTeX - Zusammenfassung

- Bereitstellung eines deutschen klinischen Text-Corpus
  - Umfangreich (60 k annotierte Dokumente)
  - Standardisiert
  - Definierte Qualität
  - Offen
- Unter Governance der MII
- 16 Partner
- Nachhaltigkeit durch Community-Integration
- Nutzung von LLM ⇒ LLM WS
- Integration der Ergebnisse in NUM