DIFUTURE
Data Integration for Future Medicine

HiGHmed
Medical Informatics

miracum
Medical Informatics in Research and Care in University Medicine

SMITH
Smart Medical Information Technology for Healthcare

GeMTeX

German Medical Text Corpus

Methodenplattform

Martin Boeker

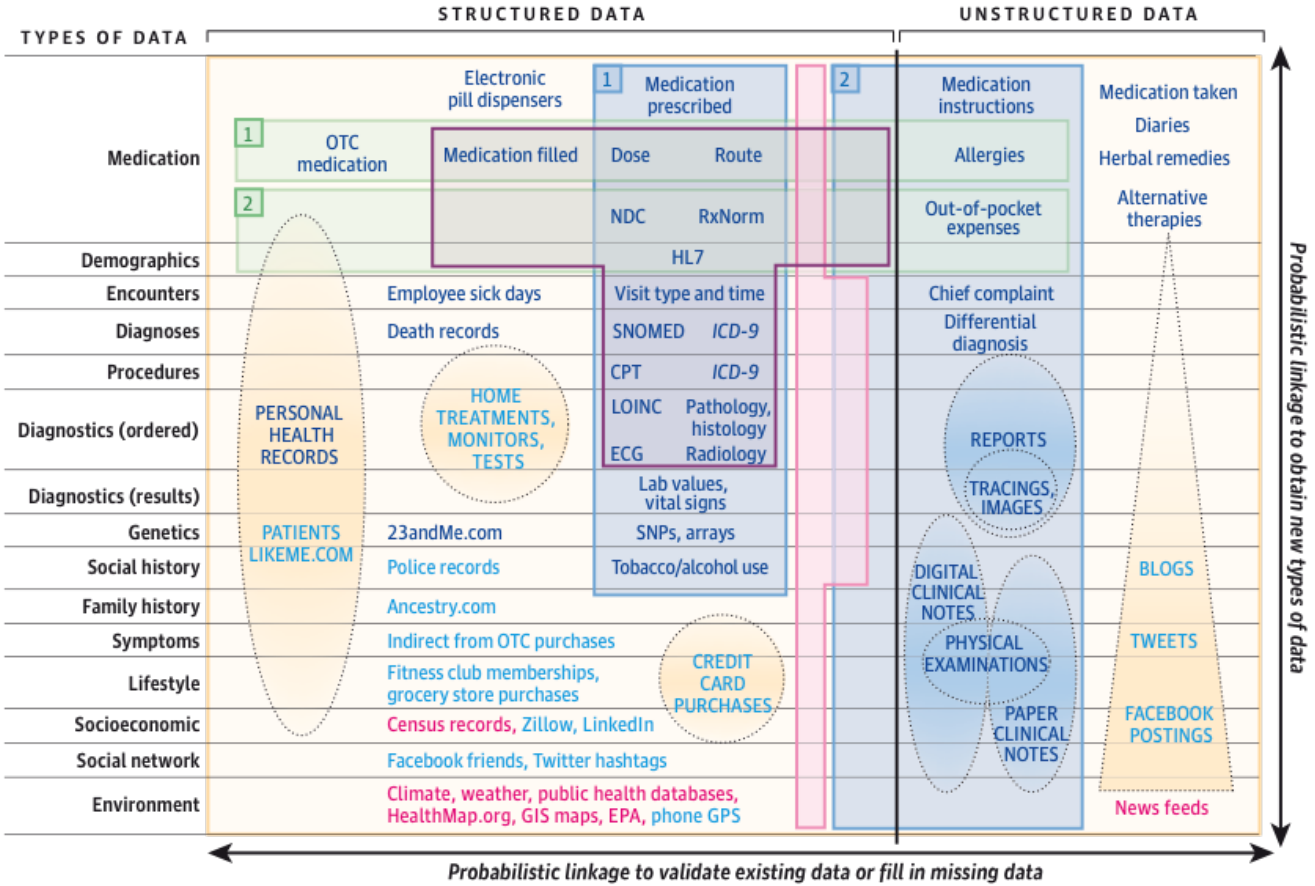
MEDIZIN INFORMATIK INITIATIVE

2023-12-13

MII-Symposium Berlin

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Strukturierte vs. unstrukturierte Daten - hoher Anteil textbasierter Daten



Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the Missing Link for Big Biomedical Data. *JAMA* (2014) doi:10.1001/jama.2014.4228.

Charakteristika klinischer Texte

Phänomen	Beispiel	Erläuterung
Telegrammstil	“Weitere Abklärung auf Intensiv”	Unvollständige Sätze, skizzenhafte, stichwortartige Ausdrucksweise
Umgangssprachlichkeit	“Coronaverdacht”, “Leberlatte”	Oft abhängig vom klinischen “Milieu”
Ad-hoc-Abkürzungen	“lymphozyteninfiltr.”	Weglassen des Wortendes mit oder ohne Punkt
Mehrdeutige Akronyme (In Kliniktexten selten definiert.)	“LCS”	“Long-Covid-Syndrom”, aber auch Bezeichnung einer Knieprothese (“low contact stress”) oder “Liquor cerebrospinalis”.
Kurzformen lokaler oder regionaler Bedeutung	“UKE” “St. n.” “EBA”	“Universitätsklinikum Hamburg-Eppendorf”; “Status nach” = “Zustand nach” (Schweiz) Interdisziplinäre Notfallambulanz in Graz (“Erstuntersuchung-Beobachtung-Aufnahme”)
Konventionalisierte Kurzformen durch externe Vorgaben	“V mors can dig V dext”	“Vulnus morsum canis digiti quinti dextri” = “Hundebisswunde am rechten Kleinfinger” (Nomenklatur der österr. Unvallversicherung)
Schreib- und Tippfehler	“Astra-Seneca-Impfung”, “Schüsselbein”, “Colonkrzinom”	Akzidentell oder systematisch (z.B. durch Nicht-Muttersprachler)
Schreibvarianten	“cervikal”, “Oesophagus”	Eindeutschungsregeln werden durch nicht oder nur teilweise beachtet
Nominalkomposita	“Außenmeniscusscheibendeformität” “Ibuprofenintoxikation”	lexikalisch nicht erfasste Langwörter durch Zusammenschreibung
Anaphern	(i) “Adeno-Ca Rectum pN+MX G2 (...). Tumor in toto exzid.” (ii) “Im Magen kein Blut (...). Zahlr.Schleimhauterosionen”	Präzise Bedeutung erschließt sich nur durch Bezugnahme auf den Vortext, in (i) meint “Tumor” das zuvor exakt beschriebene Karzinom; in (ii) sind “Schleimhauterosionen” “Erosionen der Magenschleimhaut”.
Negationen	“Kein Anhalt für Pneumonie”, “Pulmones: nihil” “metastasenfrei”	Oft jargontypische Wendungen
Unsicherheit, Verdacht	“V.a. Myokardinfarkt DD Lungenembolie”	Durch Kürzel wie “V.a.” (“Verdacht auf”) oder DD (“Differentialdiagnose”) ausgedrückt
Zeitbezüge	“Z.n. Covid-19”, “Streptokokkenangina 06/16”	“Z.n.” (Zustand nach) verweist auf früheres Ereignis, häufig Zeitangaben im Format “MM/YY”
Sonstige Kontexte	(i) “Vater: Pankreas-Ca” (ii) “Von Bauchlagerung wurde Abstand genommen”	(i) Familienanamnese bezieht sich auf Erkrankungen von Angehörigen. (ii) nicht ausgeführte Planungen

Stefan Schulz. Maschinelle Analyse klinischer Sprache in Zeiten von COVID-19. mdi 4/2021, 107–111 (2021).

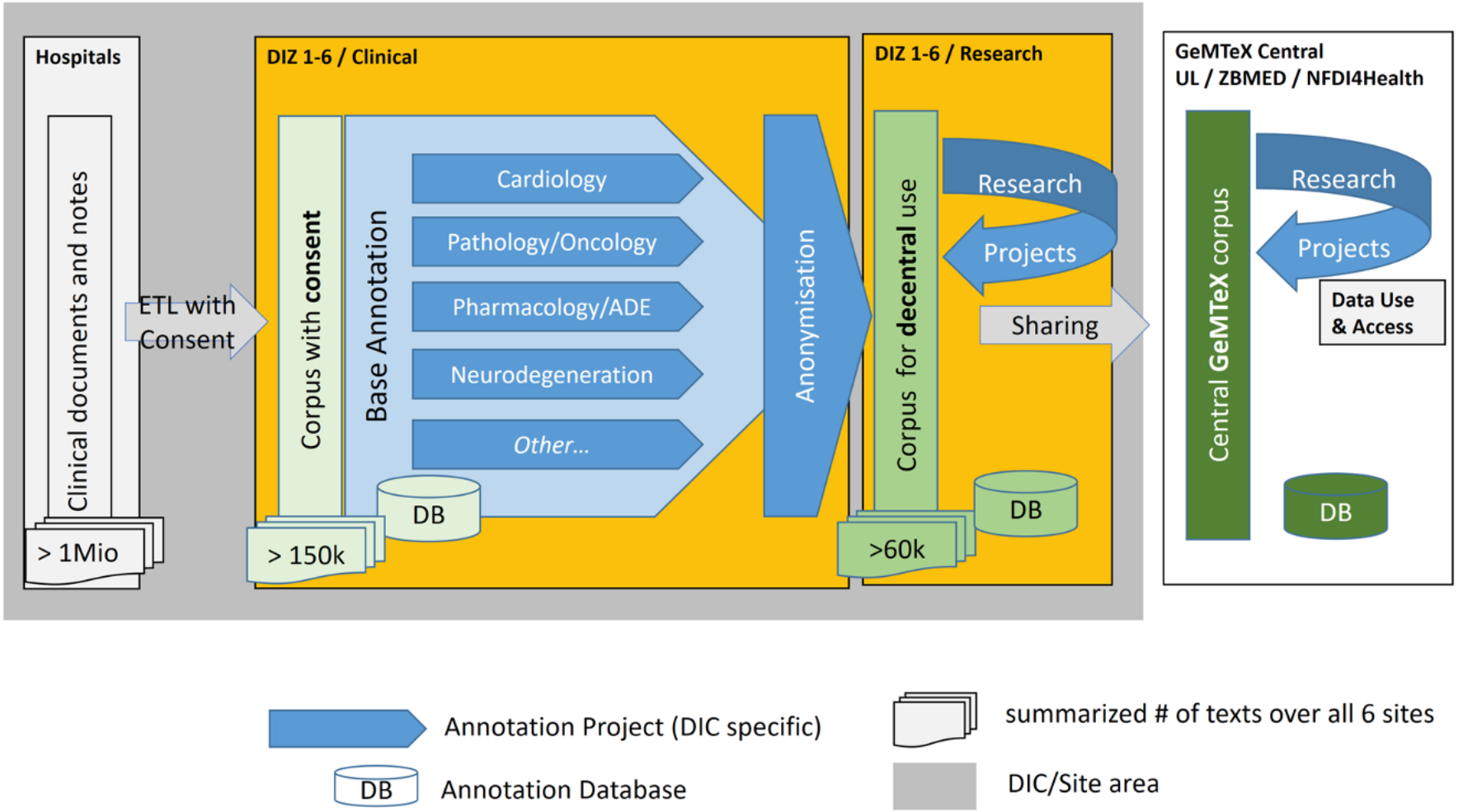
Zielsetzung GeMTeX

- Deutsches 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 DIZ
- Initialisierung von Folgevorhaben
  - Demonstration von Vorteilen der semantischen Textanalyse für die Krankenversorgung

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

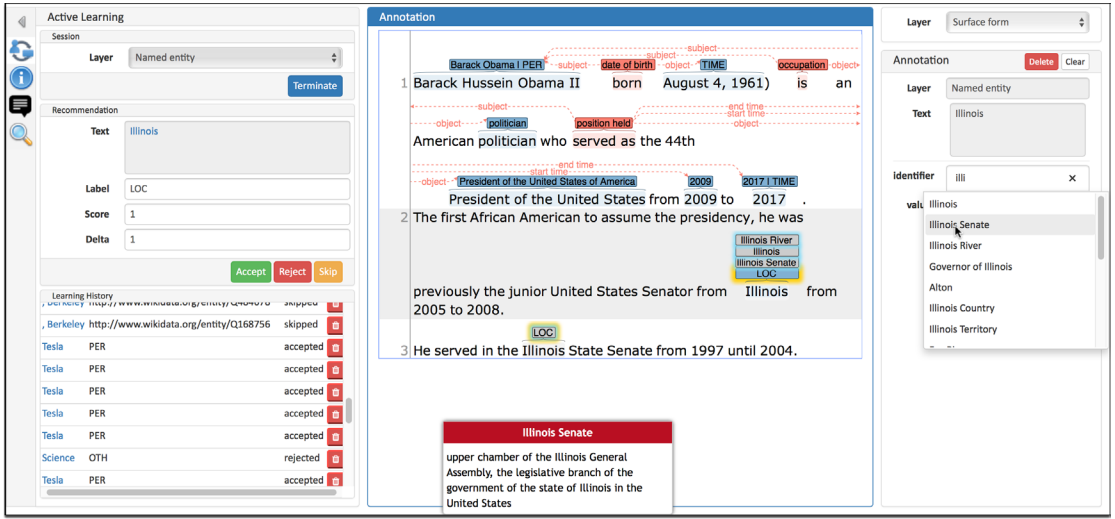

Hospitals
Clinical documents and notes
> 1Mio
ETL with Consent
DIZ 1-6 / Clinical
Corpus with consent
Base Annotation
Cardiology
Pathology/Oncology
Pharmacology/ADE
Neurodegeneration
Other…
DB
> 150k
Anonymisation
DIZ 1-6 / Research
Corpus for decentral use
Research
Projects
Sharing
DB
>60k
GeMTeX Central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

Averbis Health Discovery

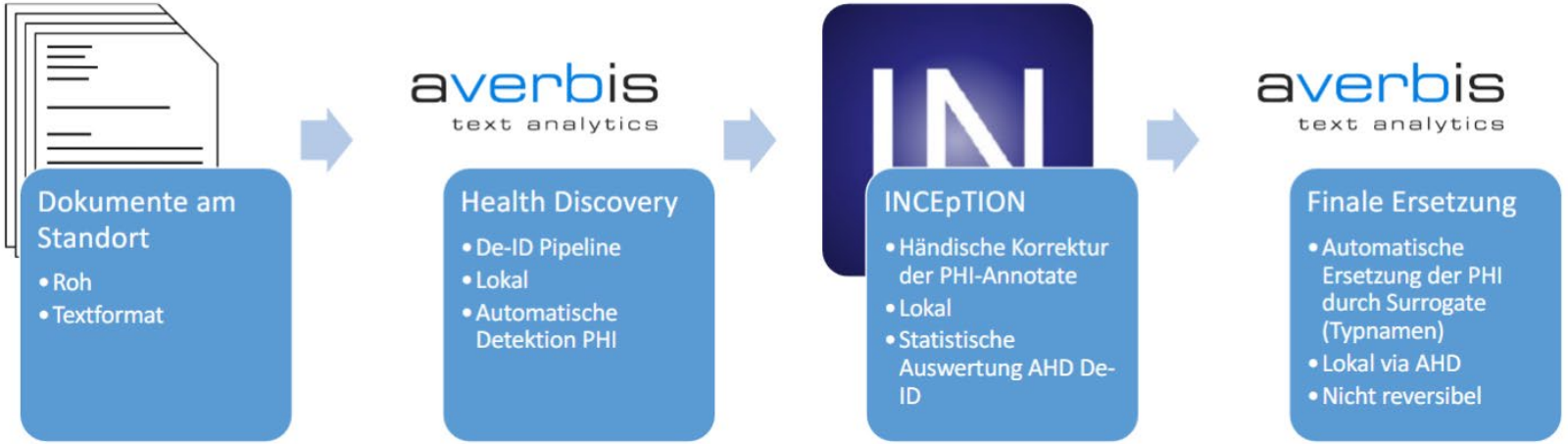
- Viele NLP Werkzeuge
  - De-Identifikation
  - Annotation
    - Diagnosen
    - Medikation
- Integrierte Workflow-Engine
  - Pipelining von Prozessschritten
  - Python-API ermöglicht Einbinden eigenen Tools
- Support durch Averbis

Annotationseditor: INCEpTION

- Kooperation mit TU Darmstadt
- https://inception-project.github.io/
- Anpassungen für GeMTeX:
  - Projektmonitoring der Annotation
  - Einbindung von Annotationsvokabularien
  - Einbindung Prä-Annotation mit interaktivem Lernen



De-Identifikationssequenz

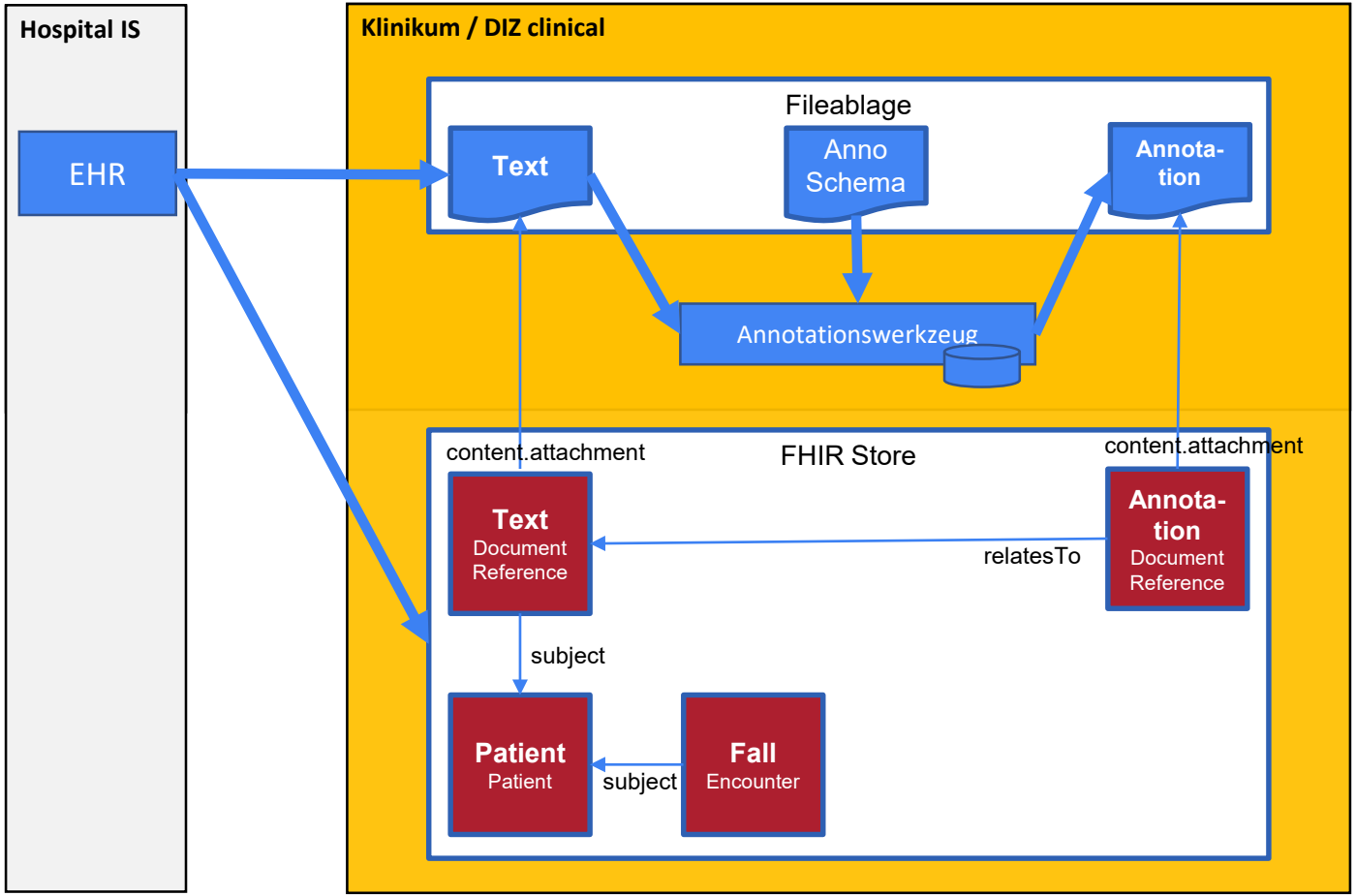


- Weitere Annotationen finden **ausschließlich** auf pseudonymisierten Texten statt
- Identifizierung der Personen **ausschließlich** über Dateinamen oder Metainformationen und Treuhandstelle
- Offset-Erhaltung **nicht** notwendig
- Ersetzung der markierten PHI-Elemente durch Typnamen (evtl. inklusive Länge des Ursprungstoken) via AHD

Annotationsmethodik

- Aufbauend auf bestehenden Methoden/Werkzeugen
  - Annotationsguideline(s) – basierend auf bestehenden GL (AIDAVA, international) & Vorarbeiten Jena
  - Annotationsterminologie(n) – SNOMED CT und LOINC
  - Automatisierte Vorannotation – AHD und spez. Projekte (HD, HPI)
  - Annotationseditor (INCEpTION)
- Annotation mit Medizinstudierenden an den Standorten
  - Mindestens Teams von 10 Studierenden erforderlich
    - Frühzeitiger Aufbau der Teams
  - Dokumentar:innen zur Ltg. der Teams/Qualitätsüberwachung
- Integration von interaktivem Lernen
- Qualitätsprüfung der Annotation

https://www.aidava.eu/



MEDIZIN INFORMATIK INITIATIVE

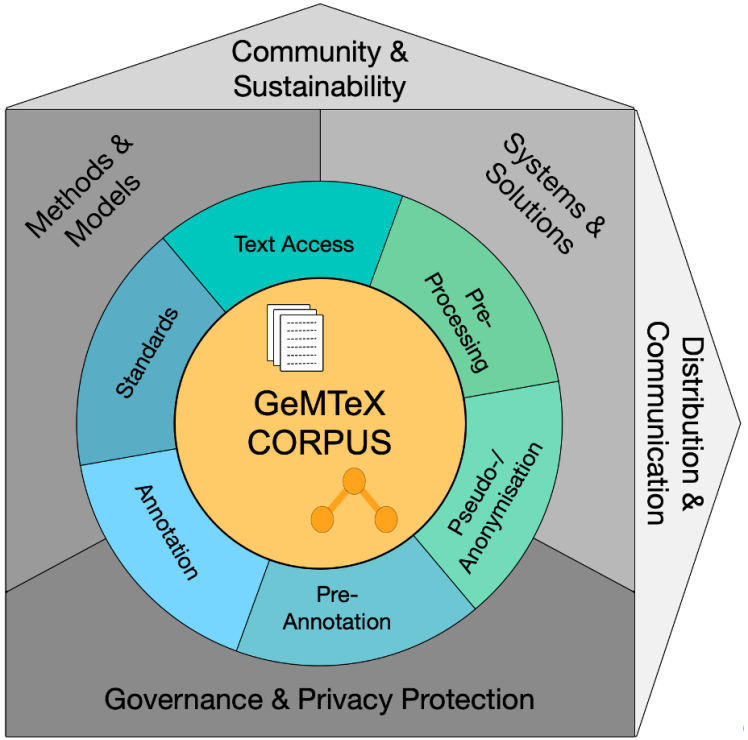
GeMTeX

Folie: Frank Meineke

– GeMTeX –

- Studienprotokoll & DS Konzept
- Annotationsguidelines
- Aufbau Technik an Standorten

- Start der Schulungen
- 01.06.2024 Start der Annotation



Vielen Dank!

MEDIZIN INFORMATIK INITIATIVE

GeMTeX

DIFUTURE
Data Integration for Future Medicine

HiGHmed Medical Informatics

miracum
Medical Informatics in Research and Care in University Medicine

SMITH
Smart Medical Information Technology for Healthcare