MEDIZIN INFORMATIK INITIATIVE

GeMTeX

GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details

Christina LOHR[a,n], Jakob FALLER[b,n], Andrea RIEDEL[b,c,n], Hung Manh NGUYEN[d,n], Markus WOLFIEN[d,e,n], Justin HOFENBITZER[f,n], Luise MODERSOHN[f,n], Jutta ROMBERG[g,n], Fabian PRASSER[g,n], Jazia OMEIRAT[h,i,n], Yutong WEN[h,i,n], Oksana GALUSCH[j,n], Udo HAHN[a,n], Marvin SEIFERLING[k,n], Christoph DIETERICH[k,n], Peter KLÜGL[l,n], Franz MATTHIES[a,n], Janina KIND[m,n], Martin BOEKER[f,n], Markus LÖFFLER[a,m,n] and Frank MEINEKE[a,n]

[a] Institute for Medical Informatics, Statistics, and Epidemiology, Leipzig University, Leipzig, Germany [b] Erlangen University Hospital, Medical Center for Information and Communication Technology, Erlangen, Germany [c] Friedrich-Alexander-Universität Erlangen-Nürnberg, Medical Informatics, Erlangen, Germany [d] Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany [e] Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden, Germany [f] Technical University of Munich, School of Medicine and Health, Institute for AI and Informatics in Medicine, TUM University Hospital, Munich, Germany [g] Data Integration Center, Berlin Institute of Health (BIH) at Charité, Berlin, Germany [h] Central IT Department, Data Integration Center, University Hospital Essen, Essen, Germany [i] Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany [j] Data Integration Center, University of Leipzig Medical Center, Leipzig, Germany [k] Klaus Tschira Institute for Integrative Computational Cardiology, University Hospital Heidelberg, Heidelberg, Germany [l] Averbis GmbH, Freiburg, Germany [m] Leipziger Forschungszentrum für Zivilisationserkrankungen – LIFE Management Cluster, Leipzig, Leipzig University, Germany [n] GeMTeX Consortium of the German Medical Informatics Initiative

10.09.2025

Federal Ministry of Research, Technology and Space

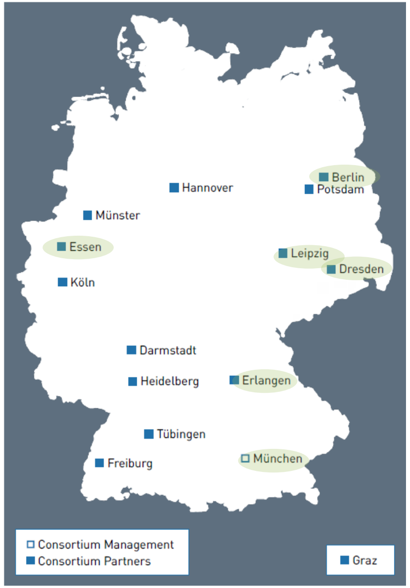
One year ago - **GMDS 2024**

- “De-Identifying GraSCCo –A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus”

- **GeMTeX**: **Ger**man **M**edical **Tex**t Corpus
  - For Training, Evaluation and Finetuning of Large Language Modes

- De-Identification and our data protection concept

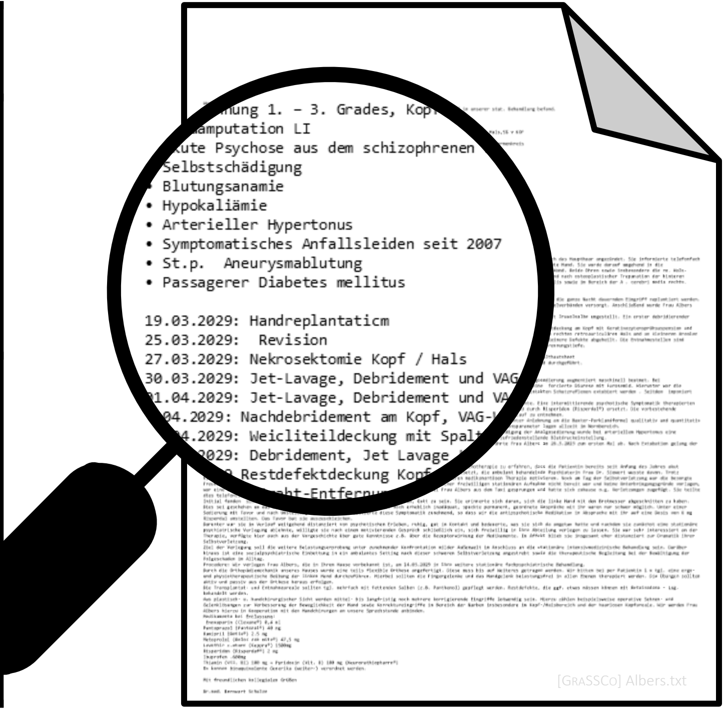
- Pilot study to answer
  - Feasability of
    - annotation tools
    - annotation workflow
    - management structures for local annotation teams
    - cross-hospital annotation requirements
    - (shared) annotation guidelines
    - appropriate ?



GRASCCO

Graz Synthetic Clinical Text Corpus

63	synthetic discharge summaries
5,430	sentences
43,667	tokens
licence	CC 0 1.0 Universal („No copyright“)
download	https://doi.org/10.5281/zenodo.6539131
more details	Luise Modersohn, et al. „GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus“. Stud Health Technol Inform. GMDS 2022, 2022 Aug 17;296:66-72.

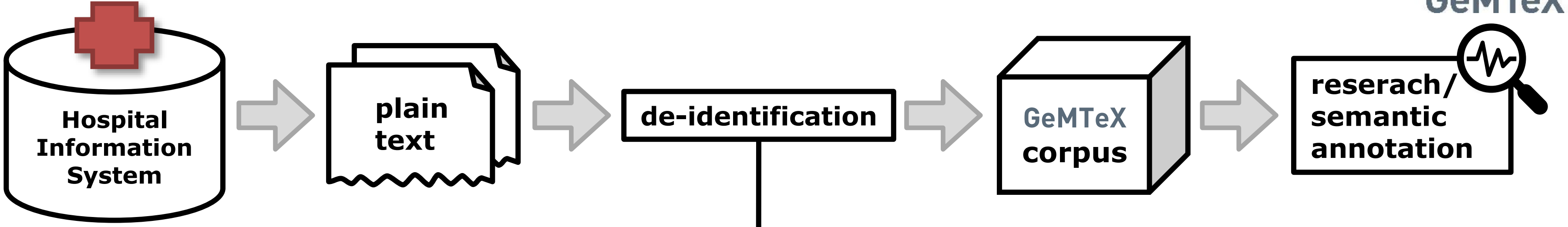


nung 1. – 3. Grades, Kop
amputation LI
kute Psychose aus dem schizophrenen
Selbstschädigung
• Blutungsanamie
• Hypokaliämie
• Arterieller Hypertonus
• Symptomatisches Anfallsleiden seit 2007
• St.p. Aneurysmablutung
• Passagerer Diabetes mellitus

19.03.2029: Handreplantaticm
25.03.2029: Revision
27.03.2029: Nekrosektomie Kopf / Hals
30.03.2029: Jet-Lavage, Debridement und VAG
01.04.2029: Jet-Lavage, Debridement und VA
04.2029: Nachdebridement am Kopf, VAG-
04.2029: Weicliteildeckung mit Spalt
2029: Debridement, Jet Lavage
Restdefektdeckung Kopf

[GRASSCO] Albers.txt

GeMTeX – De-Identification



1) PII detection

[NAME_PATIENT]
Wir berichten über Ihre Patientin Beate Albers
[DATE] [DATE] [DATE]
(* 4.4.1997), die sich vom 19.3. bis zum 7.5.2029 in unserer stat. Behandlung befand.

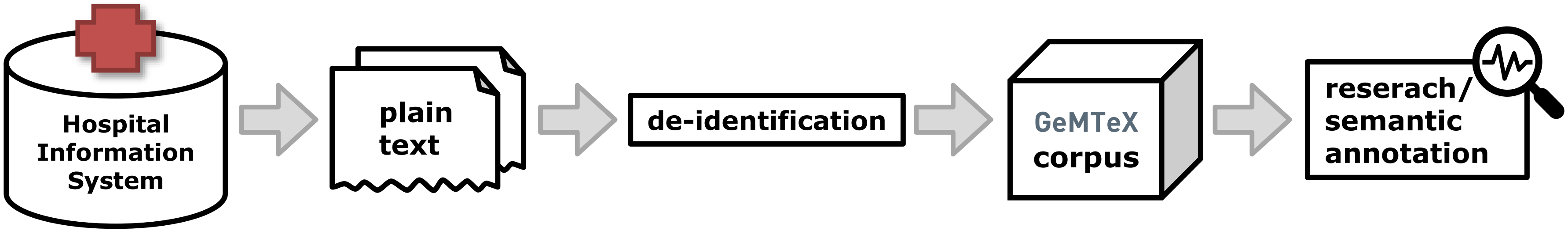
[NAME_PATIENT]
We report on your patient Beate Albers
[DATE]
(* 1997/04/04) who underwent inpatient treatment
[DATE] [DATE]
03/19 to 2029/05/07.

2) Surrogate replacement

[NAME_PATIENT]
Wir berichten über Ihre Patientin Tina Schmidt
[DATE] [DATE] [DATE]
(* 3.7.1997), die sich vom 17.6. bis zum 5.8.2029 in unserer stat. Behandlung befand.

[NAME_PATIENT]
We report on your patient Tina Schmidt
[DATE]
(* 1997/07/03) who underwent inpatient treatment
[DATE] [DATE]
06/17 to 2029/08/05.

GeMTeX – De-Identification



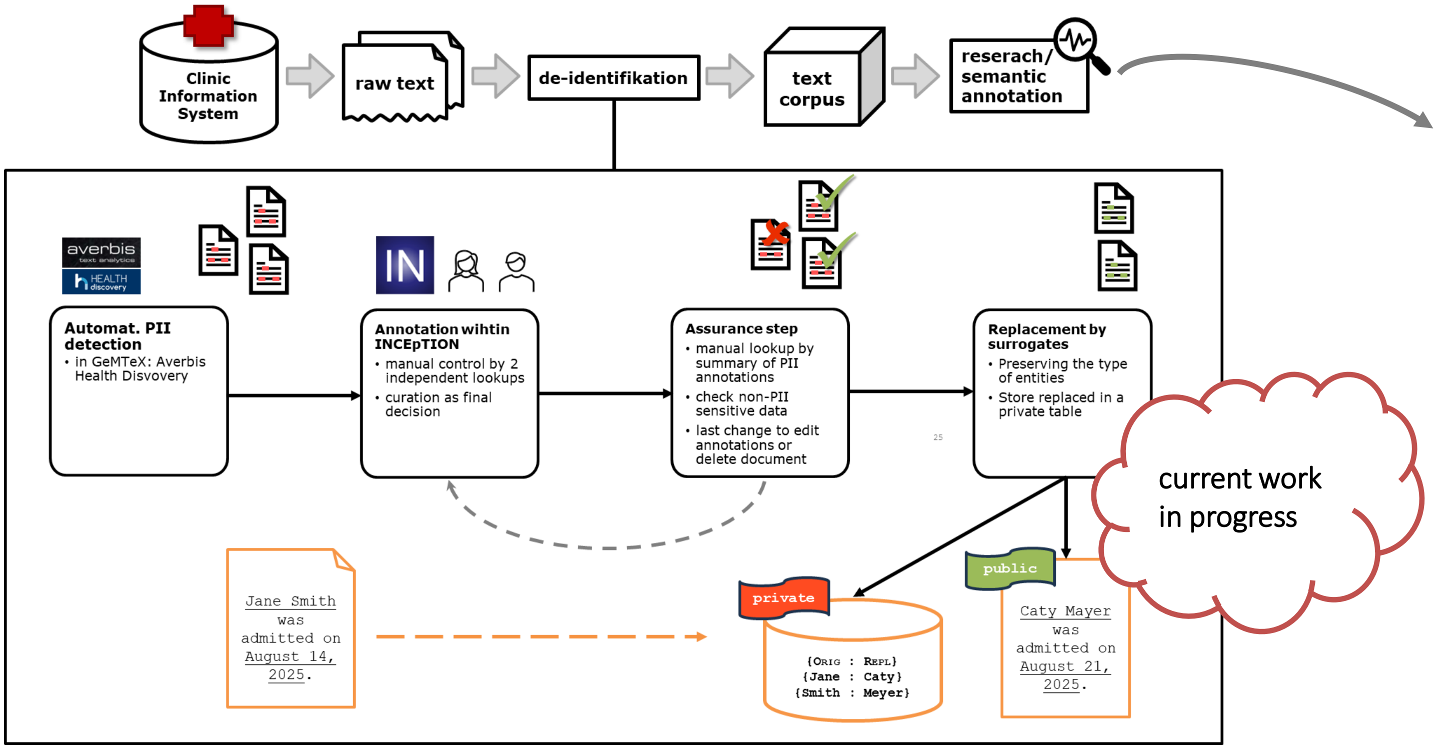
Personally Identifiable Information (PII) concept adapted from US law (PHI, HIPAA)

1. Person NAMES
2. **DATE** →
3. AGE
4. LOCATION
5. IDE
6. CONTACT
7. PROFESSION
8. OTHER

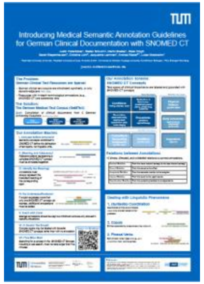
Extension:

- DATE
- DATE_BIRTH
- DATE_DEATH

GeMTeX's De-Identification in Action



Clinic Information System
raw text
de-identifikation
text corpus
reserach/ semantic annotation
averbis text analytics
HEALTH discovery
IN
Automat. PII detection
• in GeMTeX: Averbis Health Disvovery
Annotation wihtin INCEpTION
• manual control by 2 independent lookups
• curation as final decision
Assurance step
• manual lookup by summary of PII annotations
• check non-PII sensitive data
• last change to edit annotations or delete document
25
Replacement by surrogates
• Preserving the type of entities
• Store replaced in a private table
current work in progress
Jane Smith was admitted on August 14, 2025.
private
{Orig : Repl}
{Jane : Caty}
{Smith : Meyer}
public
Caty Mayer was admitted on August 21, 2025.



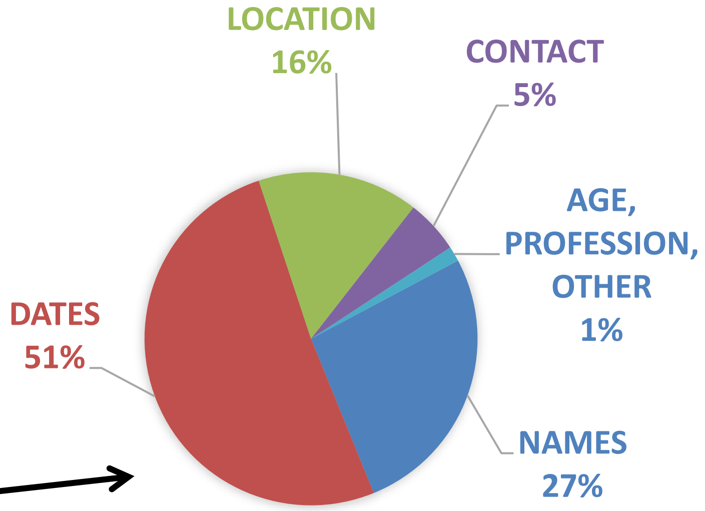
TUM
Introducing Medical Semantic Annotation Guidelines for German Clinical Documentation with SNOMED CT

Hofenbitzer, J.et al (2025).
Allgemeine Semantische Annotation – GeMTeX Annotationsrichtlinie (2.2.0). Zenodo. https://doi.org/10.5281/zenodo.15689931

Status of Corpus

GeMTeX (06/2025)

9,009	deid documents
19,475,024	alphanumeric tokens
2,162	tokens of 1 average document
377,632	PII annotations
1.94 %	token-wise ratio of PII ann. per doc



Devil’s Details

(1) A common assumption

- De-identification is done quickly if using appropriate software.
- Do not underestimate manual control steps!

Devil’s Details

(2) PII definition is vaguely regulated in Germany/EU.

- GeMTeX’s technical deid based on an industry solution, underly an US regulation and the MII data protection concept in addition to the EU law.

- Our focus: more categories than we needed

  (e.g., PROFESSION, AGE, DATES excluding BIRTH and DEATH).

Devil’s Details

(3) Details in annotations & nested entities

- locations & names in institutions,
  - „ Universitätsklinikum *Carl Gustav Carus Dresden* “
- dates vs. information about lymph nodes
  - „ … UICC 2009: pT-3c, G-3, L-1, V-1, pN-2b (7/15), pM-1 (PER) …“

Lessons Learne☝

(1) Annotation Workflow

- Take care of the **staff**!
- Plan **a training phase: 2-3 months** if the team is starting with manual annotation as a new task.
  - Shorten this step for subsequently hired staff.
- Have **frequent meetings** to **collect** and **discuss annotation questions.**
- **Be careful with conceptual changes!**

Lessons Learne☝

(2) In-house Data

… availability from project start!

- Detailed **in-house documents** for test runs should be available from the **start of the project**!
- Use **in-house lists of clinical institutions**’ names to extend the automated pipeline!

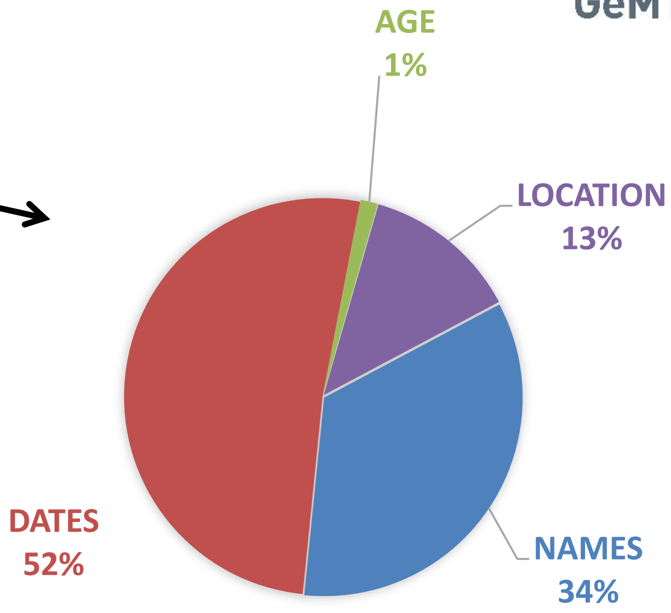
Lessons Learne☝

(3) Simplify Manual Annotation

- Examine the **requirements** in **subsequent steps**.
  - Typical PII categories may also include biometric data.
- Include **Date normalization** in de-identification.
  - … if dates are needed later.
- **Reduce** ambiguity and **cognitive** load.
  - Prioritize a simple annotation scheme, DOCTOR_NAME for medical staff.

Results

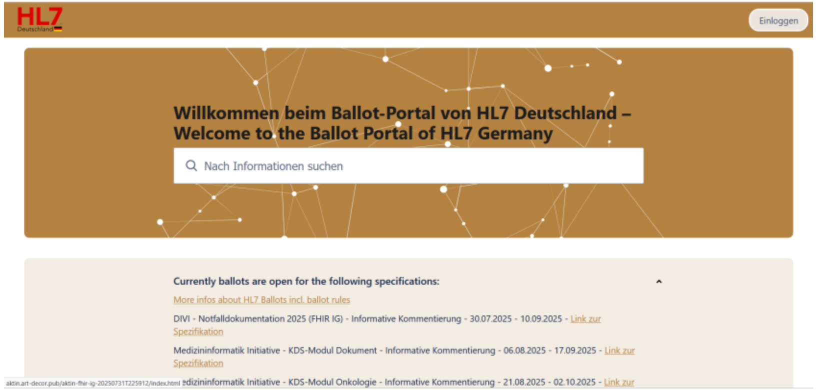
- Update GraSCCo annotations
  - 1,436 PII annotations
  - ≈ 3 % of tokens
  - provided as UIMA CAS XMI and JSON
- Update Annotation Guideline with real world examples



https://doi.org/10.5281/zenodo.15747389

Outlook

- Deidentification nearly finished
- Training of semantic annotation started with experience from deidentification
- Kerndatensatz-Modul „Dokument“ in ballot process
  - ➢ https://hl7germany.atlassian.net/servicedesk/customer/portals
  - for publication and storage
  - open until 17.09.2025
  - online feedback meeting: 12.09.2025 9:00



Willkommen beim Ballot-Portal von HL7 Deutschland – Welcome to the Ballot Portal of HL7 Germany

Nach Informationen suchen

Currently ballots are open for the following specifications:

More infos about HL7 Ballots incl. ballot rules

DIVI - Notfalldokumentation 2025 (FHIR IG) - Informative Kommentierung - 30.07.2025 - 10.09.2025 - Link zur Spezifikation

Medizininformatik Initiative - KDS-Modul Dokument - Informative Kommentierung - 06.08.2025 - 17.09.2025 - Link zur Spezifikation

aktin.art-decor.pub/aktin-fhir-ig-20250731T225912/index.html edizininformatik Initiative - KDS-Modul Onkologie - Informative Kommentierung - 21.08.2025 - 02.10.2025 - Link zur

MEDIZIN INFORMATIK INITIATIVE

GeMTeX

GeMTeX's De-Identification in Action: Lessons Learne☝ & Devil's Details

Thanks to

- **all annotators & curators**
- **DIZ & GeMTeX staff**
- **INCEpTION team and Averbis**

No conflict of interest.

Kindly contact me:

- ➢ **christina.lohr@imise.uni-leipzig.de**

This work was supported by BMFTR (previously BMBF) within the GeMTeX project under grants 01ZZ2314B (CL, FMat, UH, OG, JK, ML, FMei), 01ZZ2314A (MB, LM, JH), 01ZZ2314E (JR, FP), 01ZZ2314F (HMN, MW), 01ZZ2314G (JF, AR) and NUM 2.0 grant 01KX2121 (AR, JF), 01ZZ2314I (MS).

10.09.2025

GMDS ERHELLT GESUNDHEIT gmds 2025 70. JAHRESTAGUNG Jena · 7.-11.9.2025

Federal Ministry of Research, Technology and Space