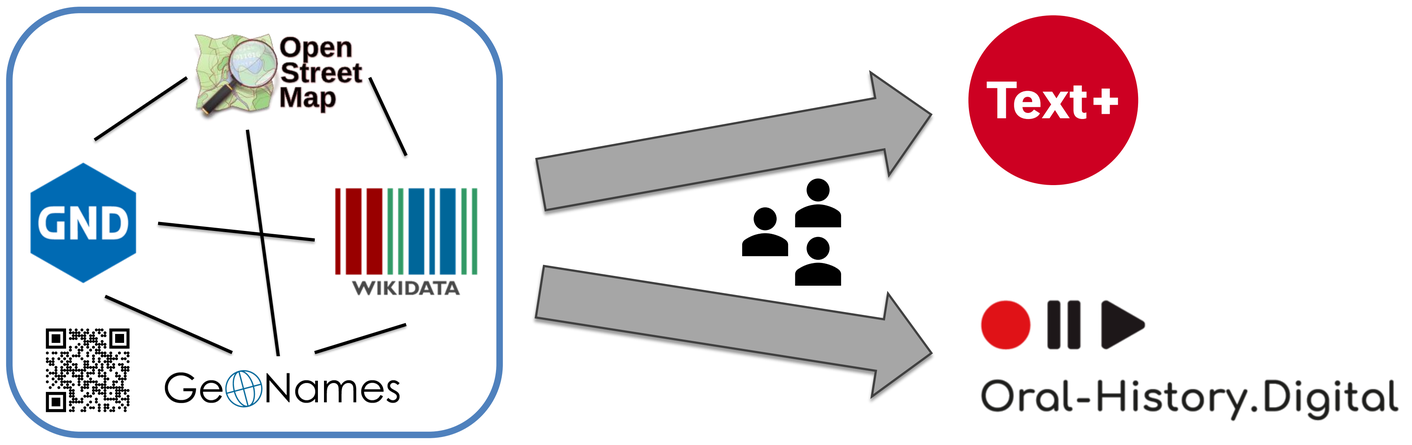
Automatisierte Typklassifikation von Normdaten mit BERT

Maximilian Hebeis, Leon Fruth, Tobias Gradl, Andreas Henrich

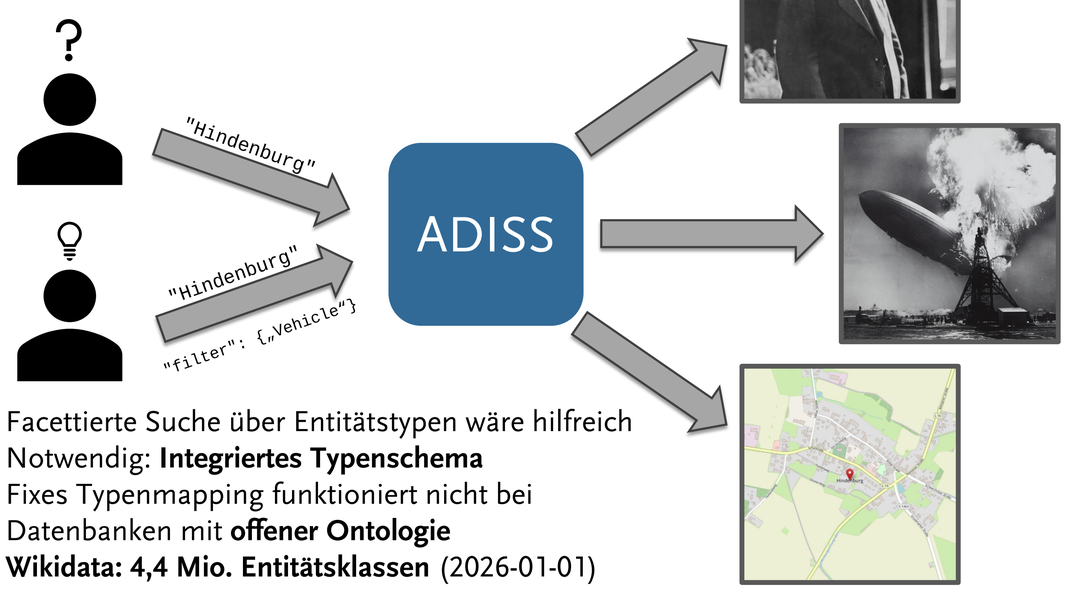
Kontext: Übergreifende Suche in Normdatenbanken

ADISS (**A**uthority **D**ata **I**ntegration **S**earch **S**ystem)

- ...verknüpft bestehende Normdaten aus verschiedenen Quellen (z. B. GND, Wikidata)
- ...überführt diese in ein integriertes Schema
- ...ermöglicht eine einheitliche Suche nach Normdaten über verschiedene Datenquellen hinweg
- ▶ ...wird von der **Text+ Registry** und **Oral-History.Digital** eingesetzt, um Forschungsdaten anzureichern



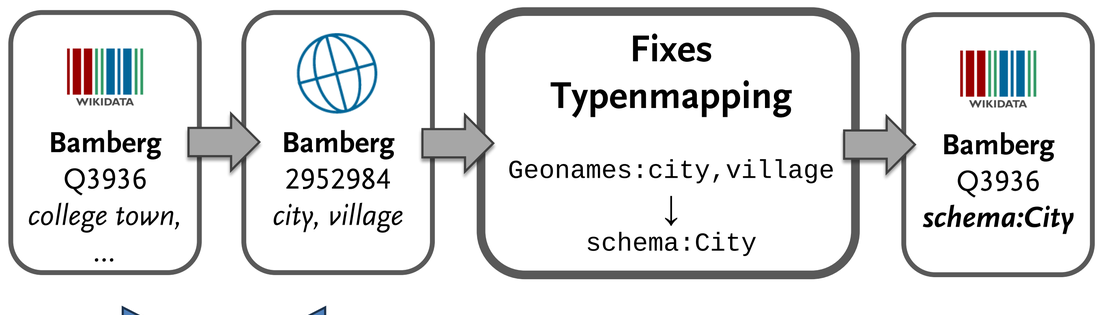
Problem



- ▶ Facettierte Suche über Entitätstypen wäre hilfreich
- ▶ Notwendig: **Integriertes Typenschema**
- ↯ Fixes Typenmapping funktioniert nicht bei Datenbanken mit **offener Ontologie**
- ↯ **Wikidata: 4,4 Mio. Entitätsklassen** (2026-01-01)

Trainingsdaten-Workflow

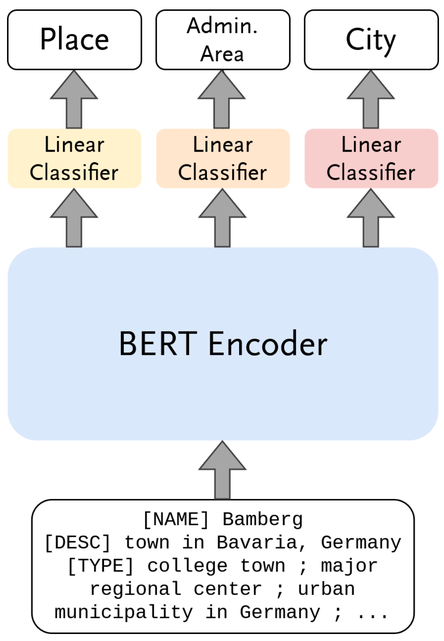
- **Idee:** Trainieren eines Klassifikators auf **verlinkten** Wikidata-Entitäten
- ▶ Für Normdaten aus Datenbanken mit rigidem Schema: Fixes Typenmapping (z.B. `gnd:AdministrativeArea` → `schema:AdministrativeArea`)
- ▶ Bestehende Verlinkungen (sameAs) von Wikidata zu GND und Geonames nutzen, um Trainingsdaten zu generieren



Modellarchitektur

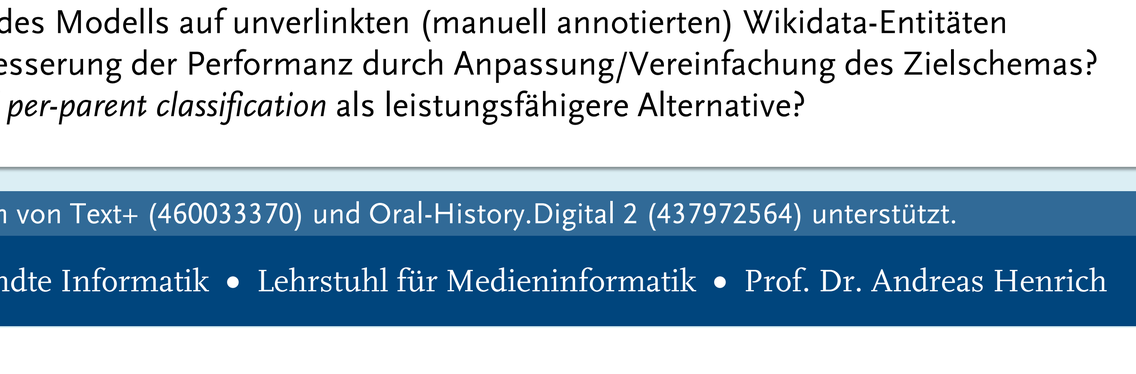
- ***Local per-level classification***
- ▶ Einbettung der Wikidata-Entität durch multilinguales **DistilBERT**-Modell
- ▶ Parallele Klassifikation der Entität für jede der drei Ebenen des Zielschemas

- **Datensatz:** Ca. 5 000 Wikidata-Items mit Ground-Truth-Label pro Zieltyp
- ▶ Training: 111 208, Test: 27 803 (Train-Test-Split von 4:1)



Zielschema

- **Schema.org**-Untermenge
- ▶ „Add-Ons“ zur möglichst genauen Abbildung relevanter Quelltypen (*PathOrBoundary*, *GroupOfPersons*)
- ▶ 33 Typen, maximale Tiefe des Hierarchiebaums von 3



Ergebnisse

Ebene	Anzahl der Klassen (*N*)	Macro Precision (%)	Macro Recall (%)	Macro F1 (%)
1	8	93,15	93.70	93,41
2	19	89,03	91,99	89,31
3	5	84,45	94,02	87,97

- **Vielversprechende Makro-F1-Werte** über alle Hierarchieebenen hinweg
- Leicht abfallende Performanz bei zunehmender Klassifikationstiefe
- Einzelne Typen performen im Vergleich schlechter (z.B. *City*, *MusicComposition*)
- ▶ False Positives sind auf jeweiliger Ebene meist als *nicht zugeordnet* klassifiziert (unter Threshold)

- **Nächste Schritte:**
- ▶ Test des Modells auf unverlinkten (manuell annotierten) Wikidata-Entitäten
- ▶ Verbesserung der Performanz durch Anpassung/Vereinfachung des Zielschemas?
- ▶ *Local per-parent classification* als leistungsfähigere Alternative?

Diese Arbeit wurde durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen von Text+ (460033370) und Oral-History.Digital 2 (437972564) unterstützt.

Otto-Friedrich-Universität Bamberg • Fakultät für Wirtschaftsinformatik und Angewandte Informatik • Lehrstuhl für Medieninformatik • Prof. Dr. Andreas Henrich