GND ; Datenqualität ; RAG

Zur Datenqualität im Rahmen eines RAG-Systems zur maschinellen Beschlagwortung

Tobias Weberndorfer

KI in Bibliotheken weiterdenken, DNB 2026

TU WIEN Bibliothek
Your knowledge hub

Inhalt

- RAG-System an der TU Wien Bibliothek
- erschlossene Daten wie und wo nutzen?
- Qualität von beschlagworteten Daten

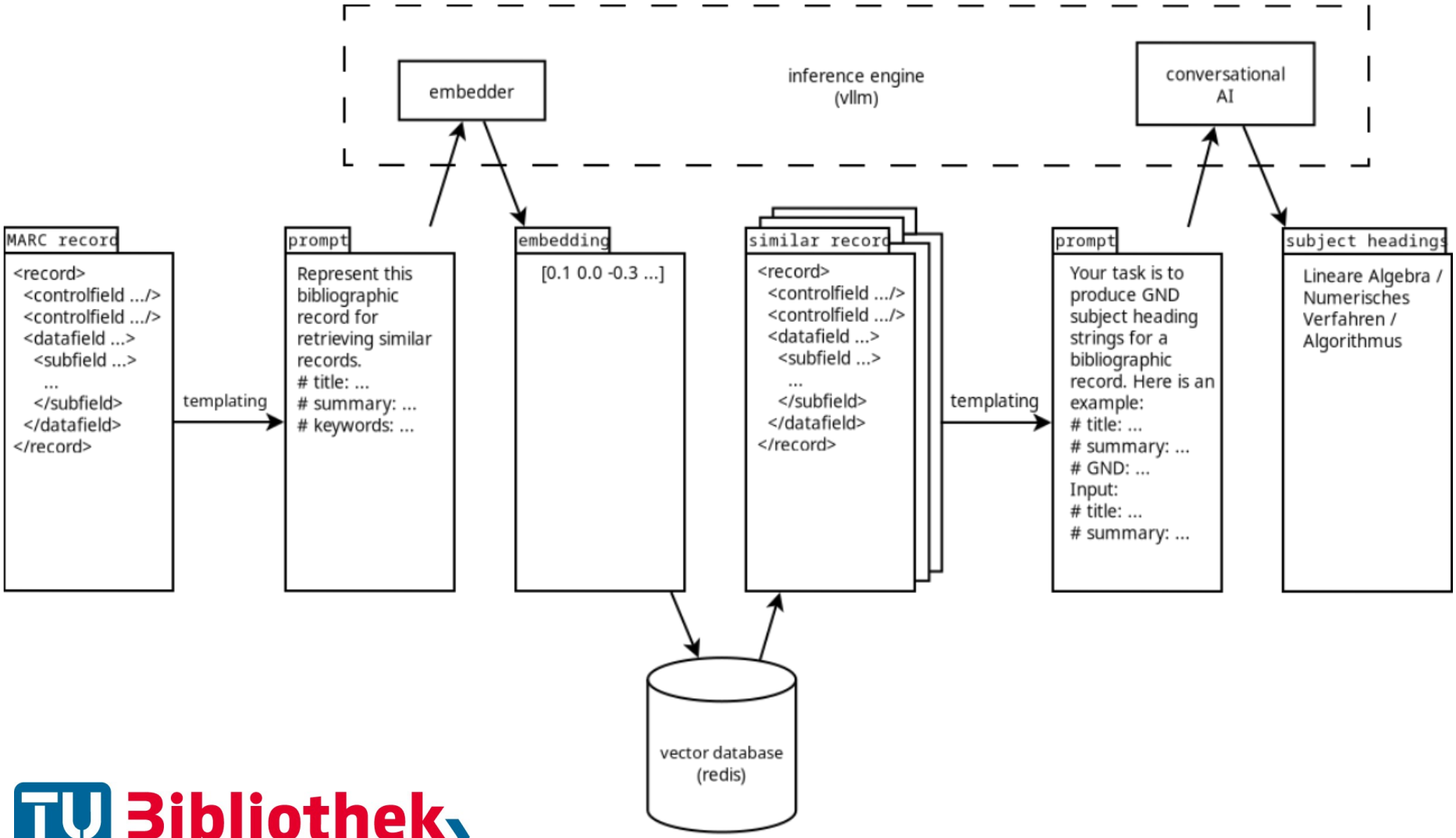
Kontext

- seit 2023 im Linked Open Data-Projekt der TU Wien Bibliothek
- Aufbau maschinelle Sacherschließung
  - intellektuelle Erschließung nicht in vollem Umfang zu leisten

Anforderungen

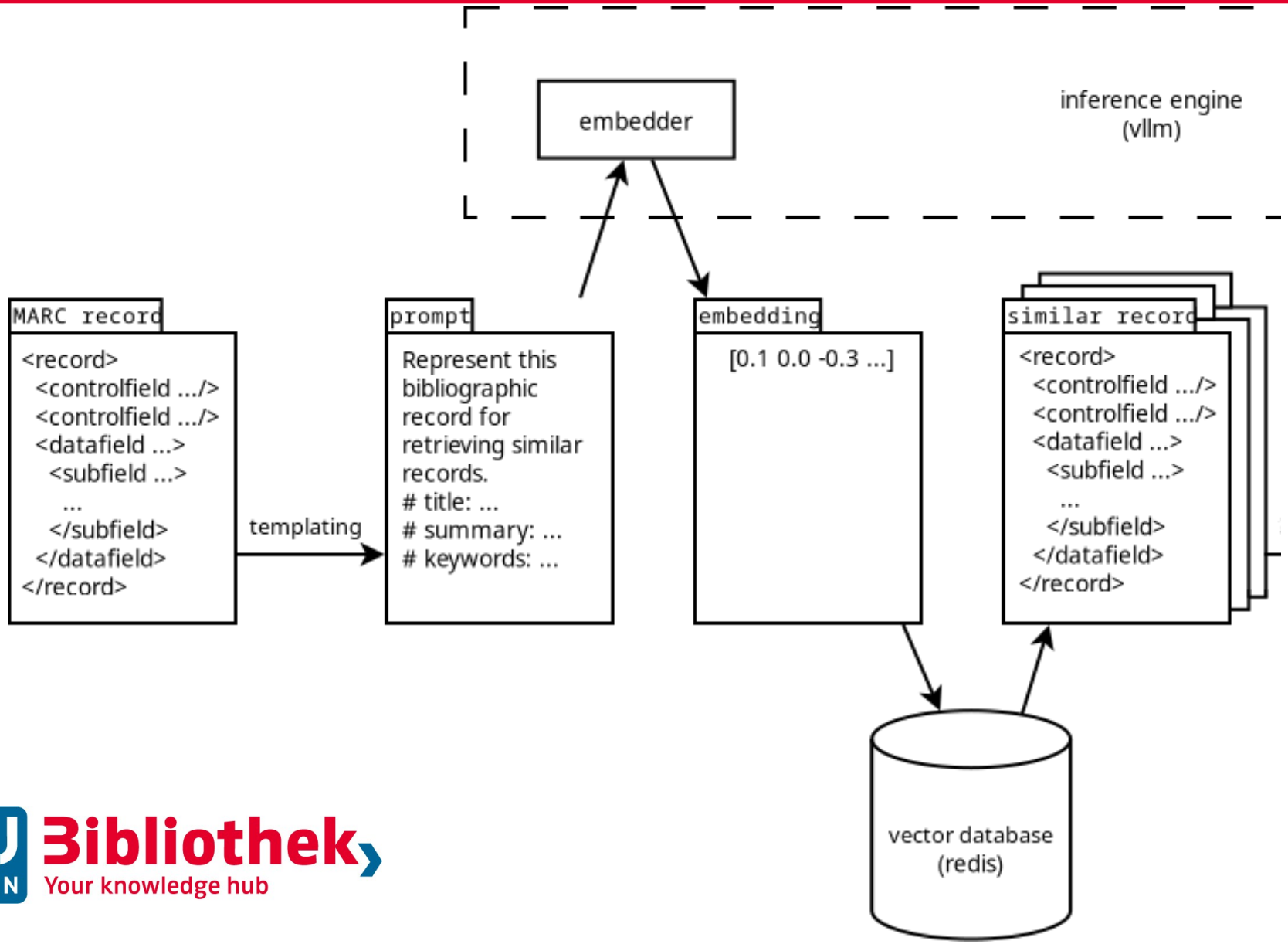
- flexibel
  - heterogene Datensätze
  - multilingual
  - multi-task: TU Wien, GND, (BK)
- unabhängig: keine externen Dienste
- ökonomisch in Bezug auf Rechenressourcen
- Qualität des Outputs
  - vergleichbar zu Übereinstimmung zwischen Erschließer*innen
  - Datenungleichgewicht!

Datenfluss (vereinfacht)



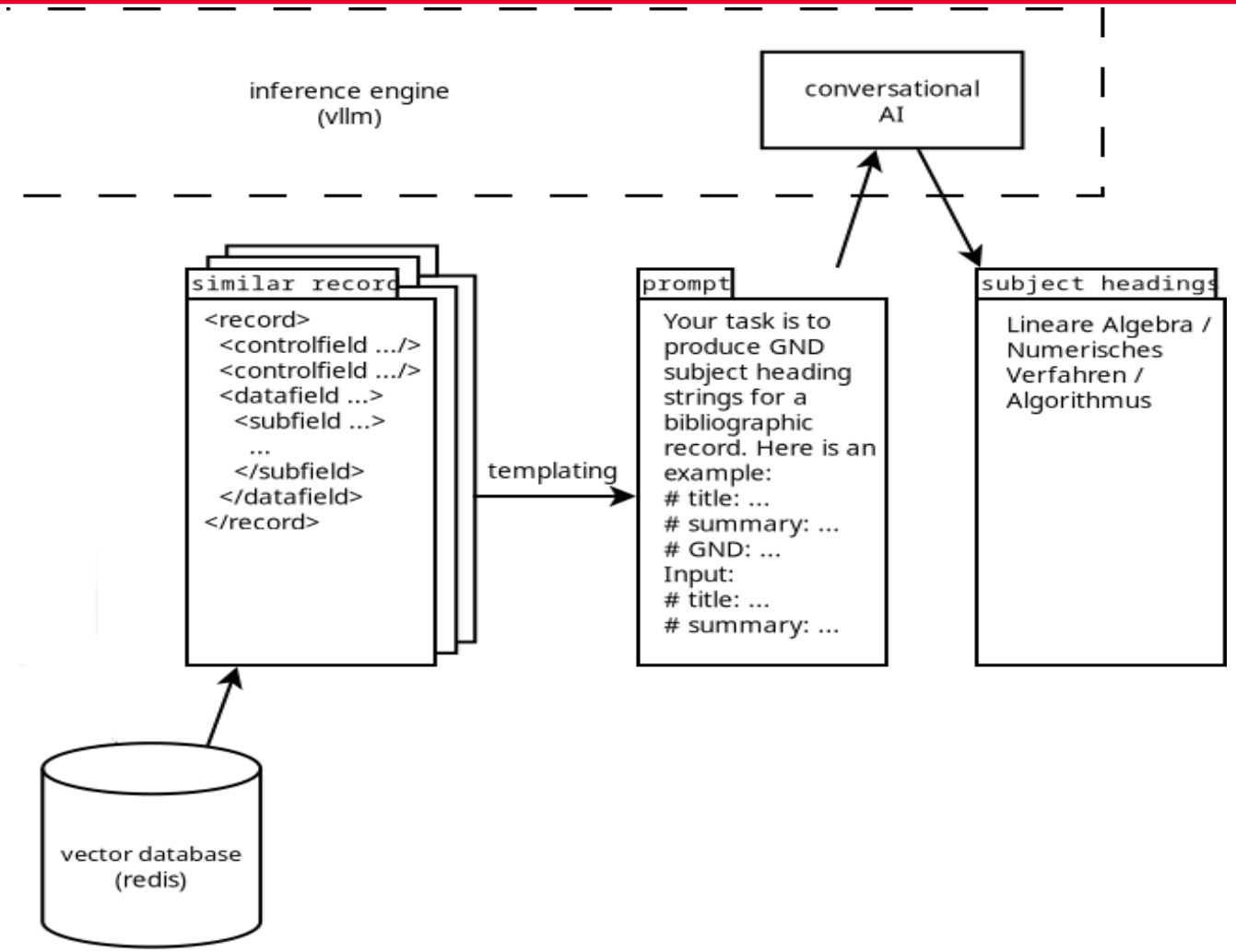
TU WIEN Bibliothek
Your knowledge hub

Datenfluss (vereinfacht)

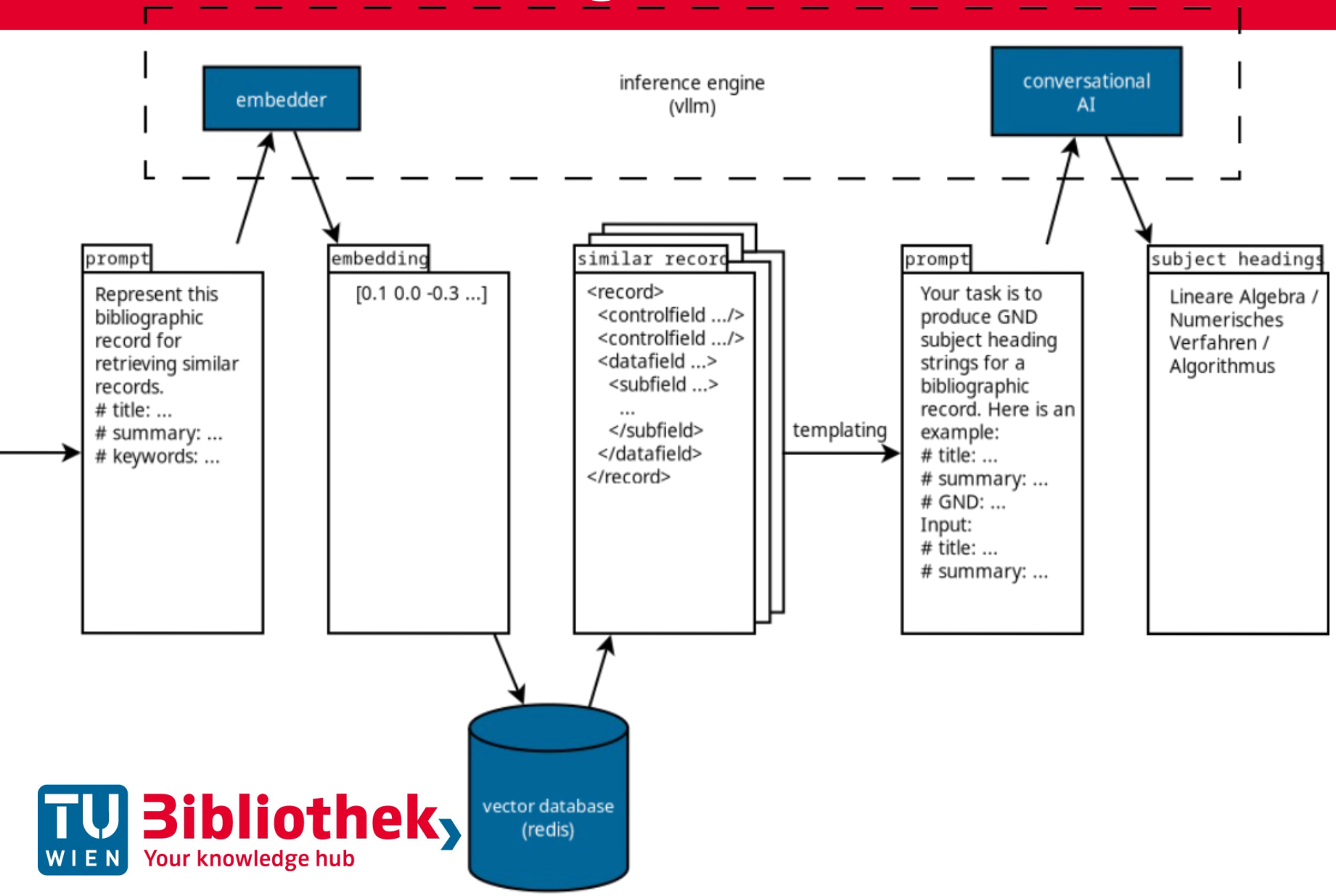


TU WIEN Bibliothek
Your knowledge hub

Datenfluss (vereinfacht)



Wo Trainingsdaten?



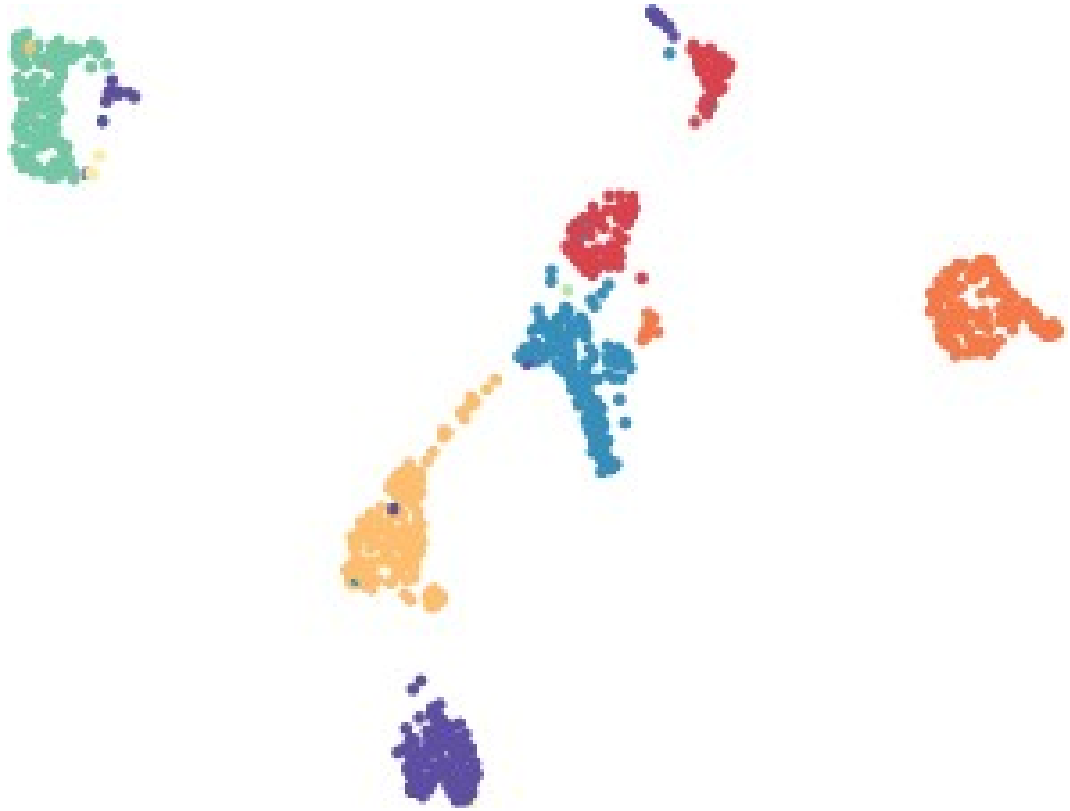
TU WIEN Bibliothek
Your knowledge hub

Retrieval

- größter Einfluss auf Leistung eines RAG-System
  1) Embedding-Modell
  2) Datenbasis Vektordatenbank
- Extrembeispiel: Treffer sehr ähnlich und tragen gewünschte Schlagworte

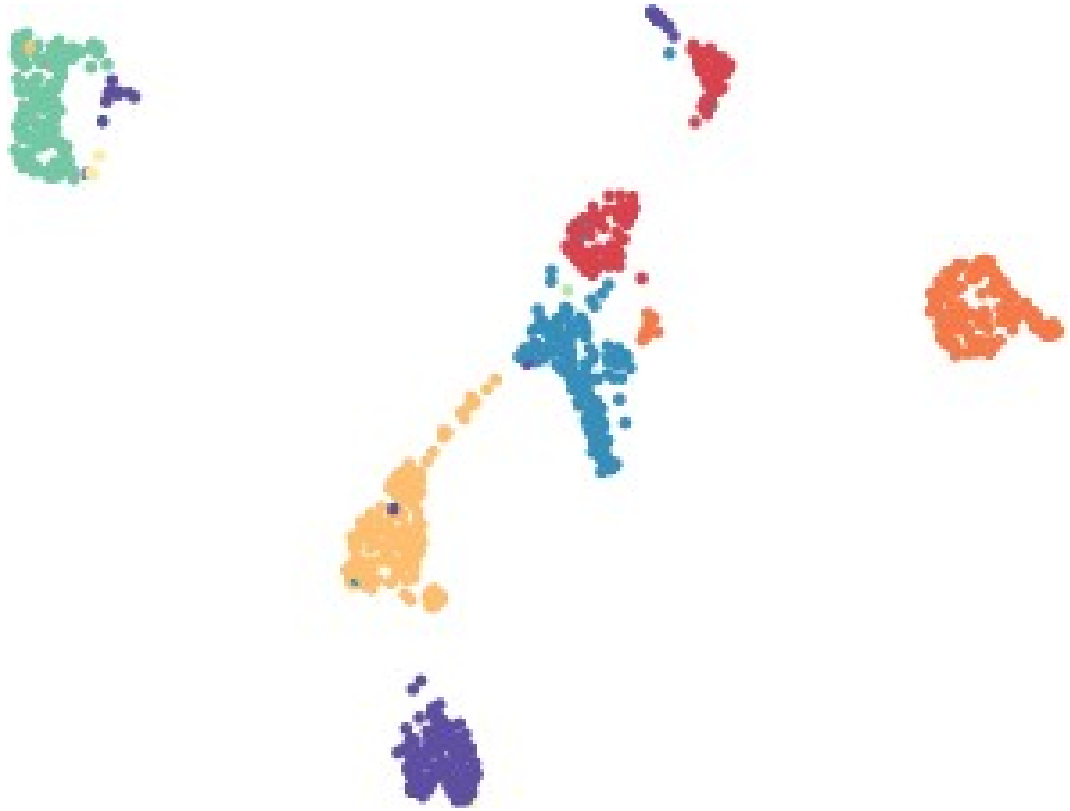
Einfluss von Noise

- Embedding-Modelle oft mit hard negative mining trainiert
  - ähnliche Dokumente mit anderen Labels



Sample selection

1) noisy Samples filtern
2) ev. noisy Samples relabelen



GND: „noisy“ Labels

- kein Goldstandard für GND
- Beschlagwortung als Interpretation
- aber...

RSWK

- Regeln für den Schlagwortkatalog
- Beispiele automatisch extrahiert
- Abgleich mit Katalog der DNB

RSWK

- RSWK

  Scientology
- DNB

  Scientology

RSWK

- RSWK

  Scientology ; Aufsatzsammlung
- DNB

  Scientology ; Aufsatzsammlung

  Esoterik ; Aufsatzsammlung

  Neue Religion ; Aufsatzsammlung

  Spiritismus

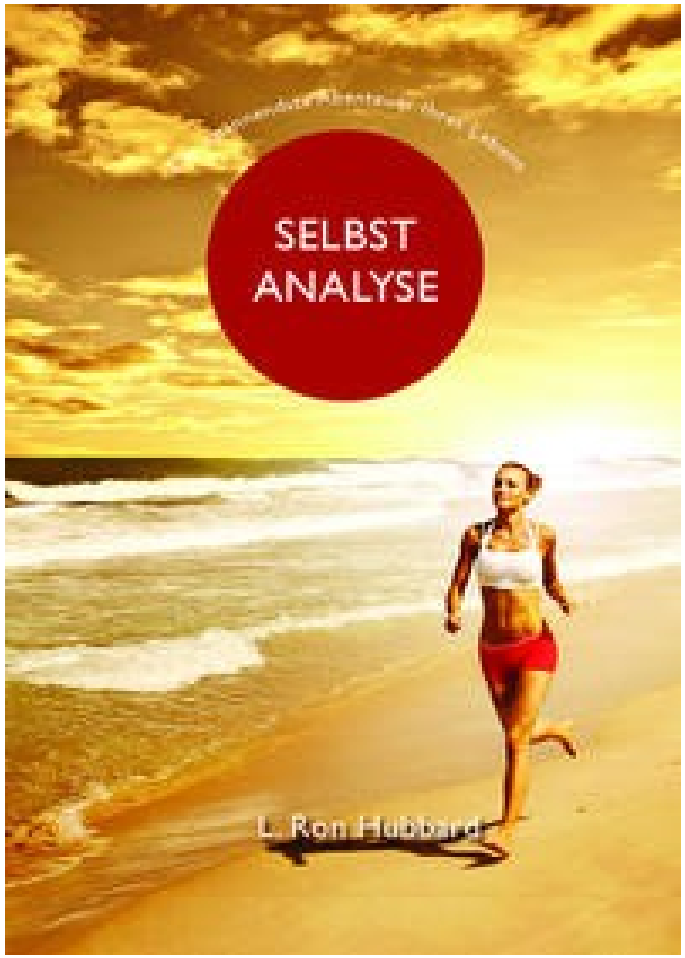
RSWK

- RSWK

  Selbstanalyse ; Dianetik
- DNB #1

  Selbstanalyse ; Dianetik
- DNB #2

  Scientology ; Lebensführung ;
  Scientology ; Test (Psychologie)



RSWK—DNB

- F1 micro: 44%
- F1 macro: 33%

Fazit

- Überblick über RAG-System zur maschinellen Beschlagwortung
- Herausforderungen von GND-Daten

VIELEN DANK!

Tobias Weberndorfer
tobias.weberndorfer@tuwien.ac.at

https://www.tuwien.at/bibliothek

TU WIEN Bibliothek
Your knowledge hub