ThWIC
Thüringer Wasser-Innovationscluster

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

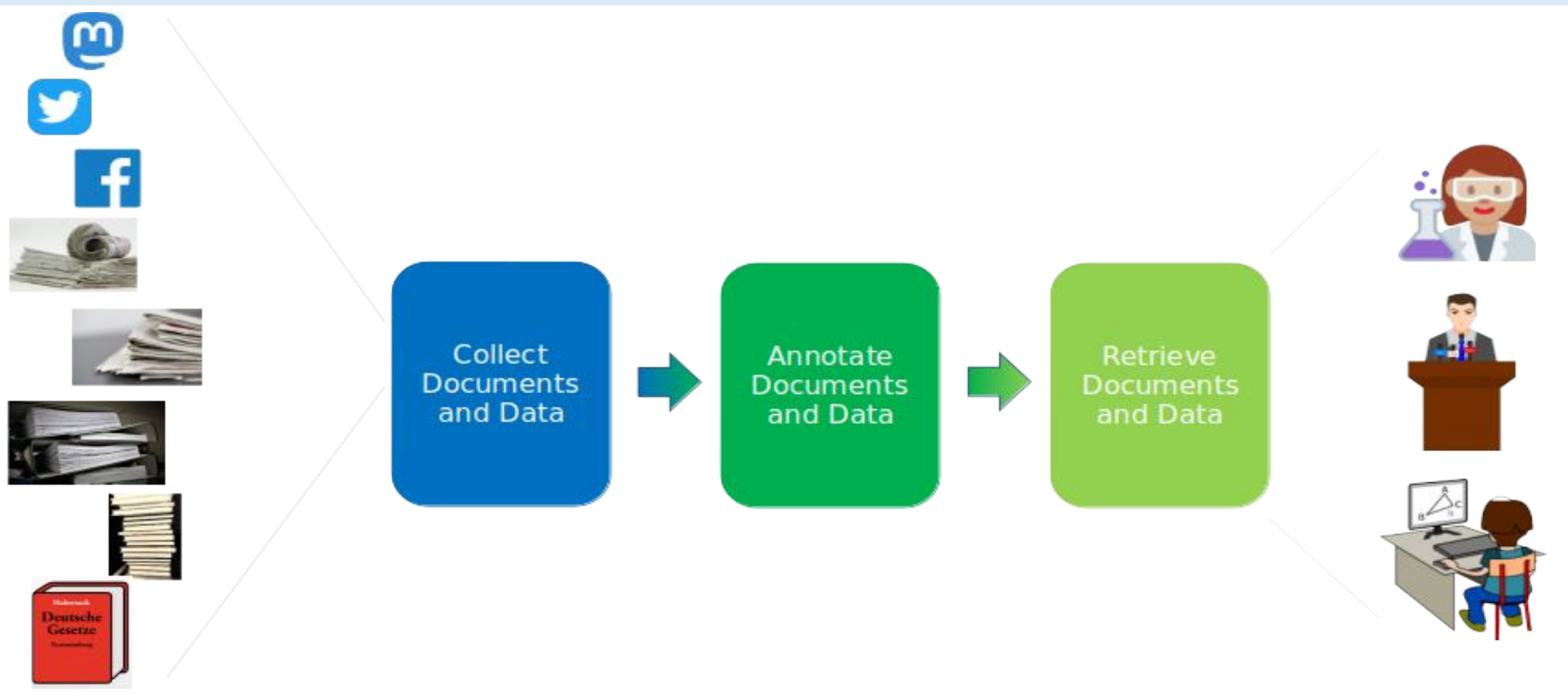
epic insights

sma DEV

ThWIC Sonar

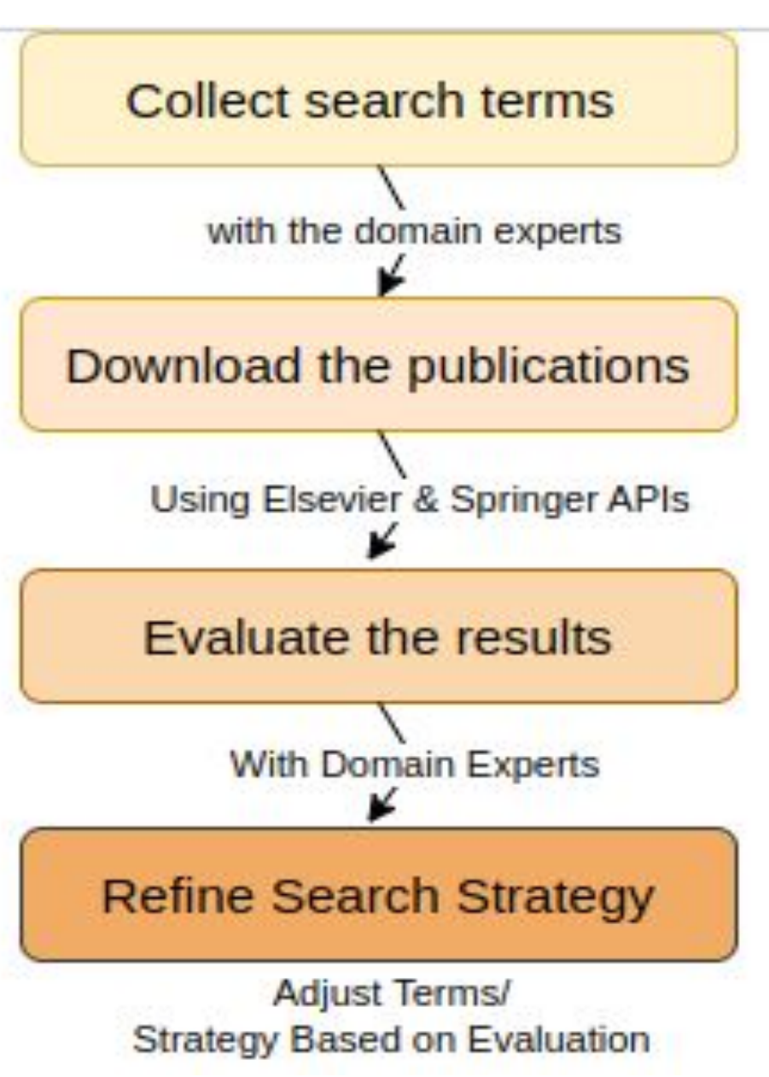
KI-basierte Navigationsunterstützung im Dokument- und Data Lake zum Thema „Wasser“

Innovations-unterstützende Maßnahmen



Collect Documents and Data
Annotate Documents and Data
Retrieve Documents and Data

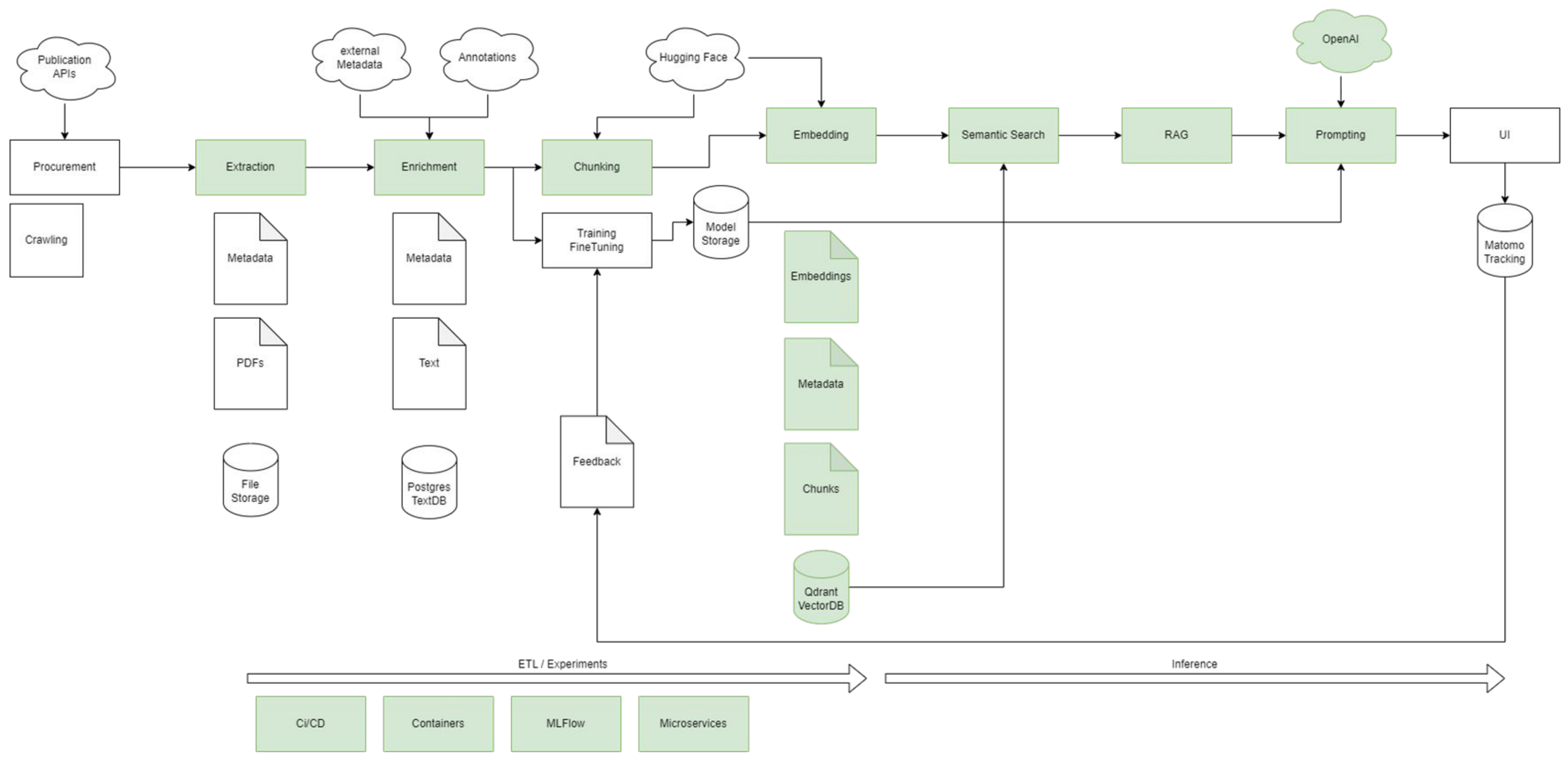
Collecting Documents: Scientific Publications & Social Media



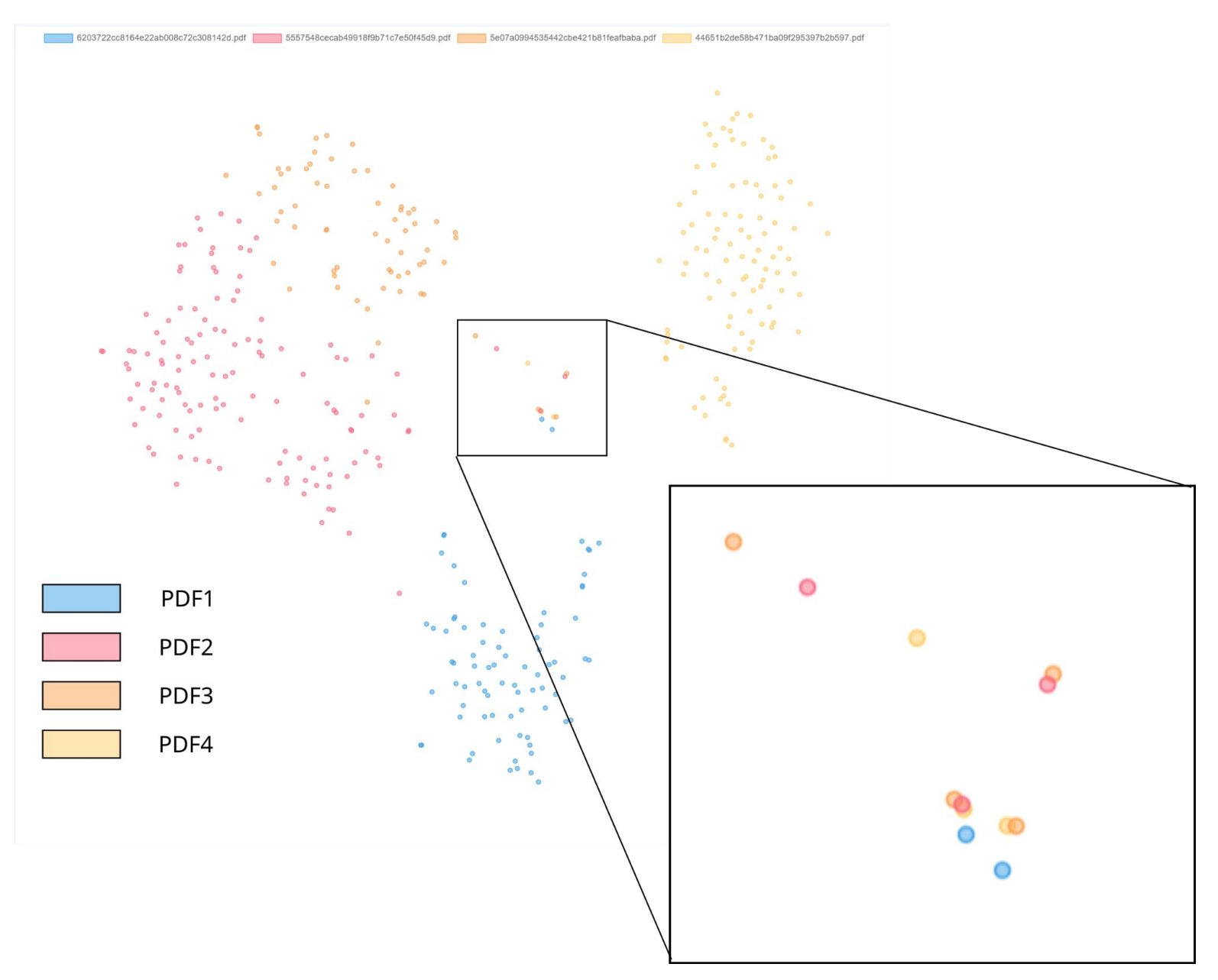
Qualification / Background of People who gave feedback in survey

1. Psychology (WaVe, Water Conflicts)
2. Literature (WaKu, American Studies)
3. Chemist (Water quality sensors)
4. Chemist (Water quality sensors)
5. Chem-informatics
6. Chemist (Water treatment)
7. Chemist (Water treatment)
8. Bioenvironmental Chemist

Current Infrastructure and System



Overview of the different components and modules of the current ThWIC Sonar AI HUB.



2D representation of vectorized PDF text fragments. Based on proximity in the vector space, we calculate similarities between text fragments and enrich the LLM message with fitting fragments. The zoomed-in box shows “References”, “License”, “Conflict of Interests”.

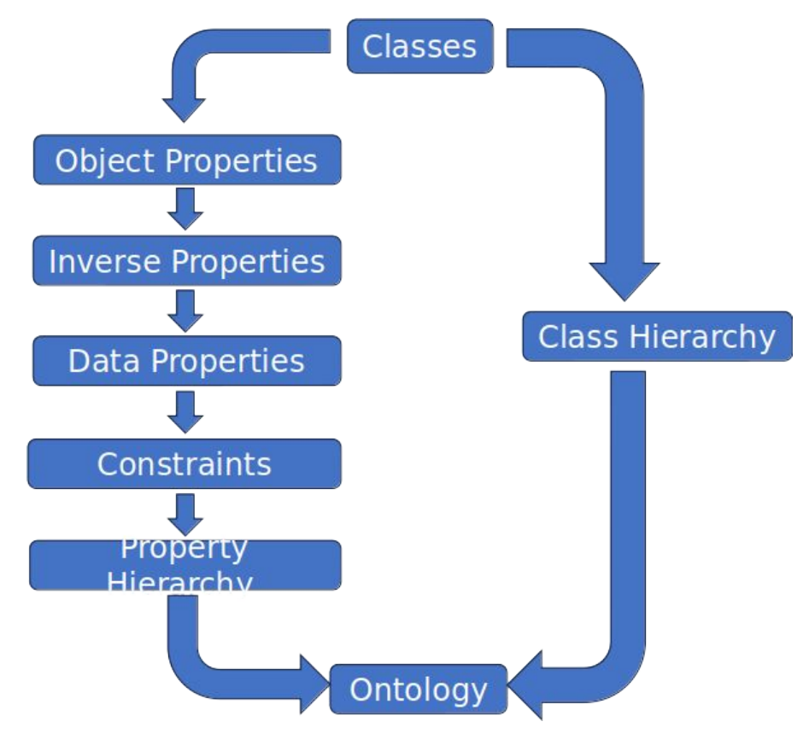
Annotating Documents

- ❑ Builds a comprehensive, high-quality dataset.
- ❑ Evaluate a subset of the data e.g., regarding relevance, or the distribution of statistical characteristics

Relevance to a given search term Is the document relevant when someone searches for ‚water purification thuringia'?	rated by experts with 0 and 1 (random samples only)
Statistical characteristics e.g., date of publication, number of citations, country of the institution	downloaded in the crawling process
Topic Classification mapping in a topical domain	classified by Neural Networks

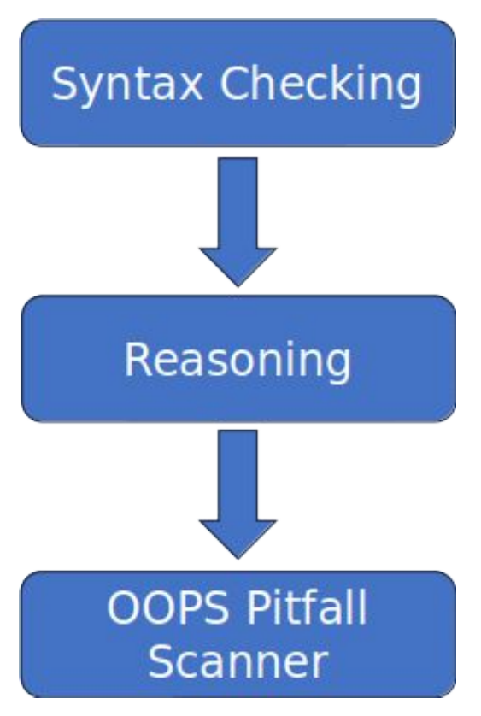
Knowledge Graph engineering with LLM's

Ontology creation : Chain-Of-Thoughts method



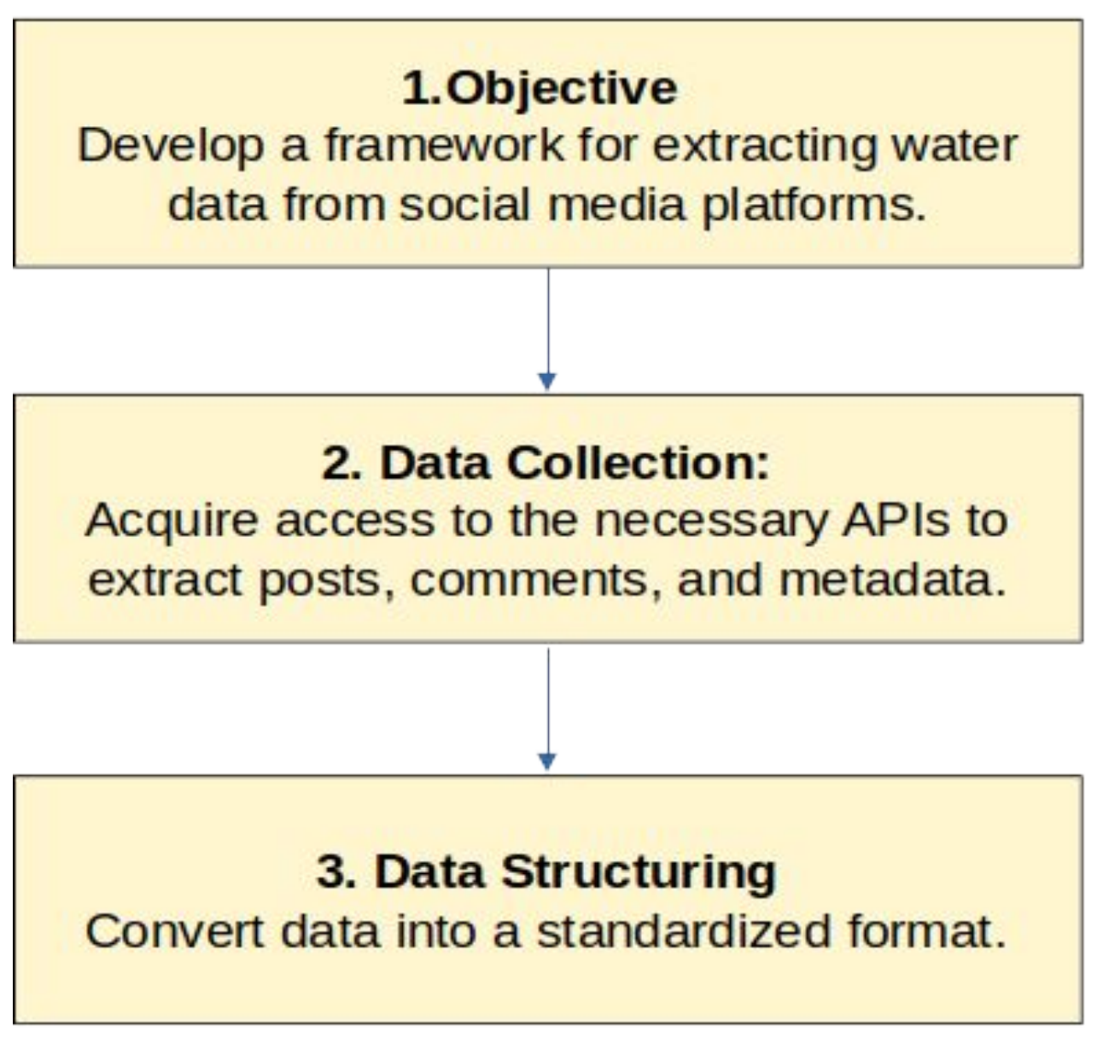
- ❑ More control over every step
- ❑ Every part can be separately modified
- ❑ Yields more detailed and qualitative results

Ontology Evaluation



- ❑ Checks the syntax of the generated ontology
- ❑ Uses the Hermit Reasoner to identify logical errors
- ❑ Checks for common mistakes of ontology creation

Current Twitter alternatives: Mastodon & BlueSky



1.Objective
Develop a framework for extracting water data from social media platforms.

2. Data Collection:
Acquire access to the necessary APIs to extract posts, comments, and metadata.

3. Data Structuring
Convert data into a standardized format.

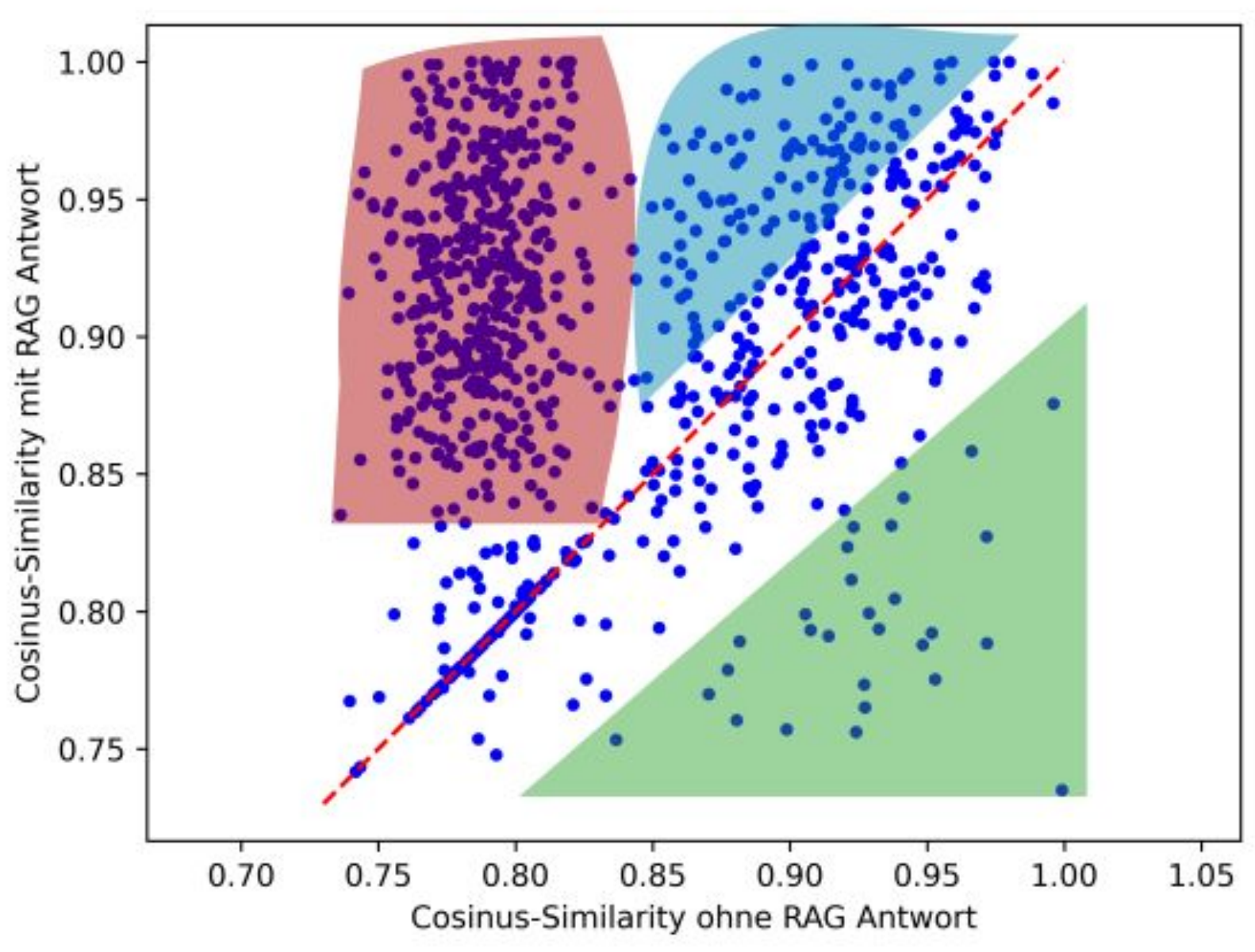
Increasing use of Social Media

- ❑ Especially in the social sciences, to get relevant data

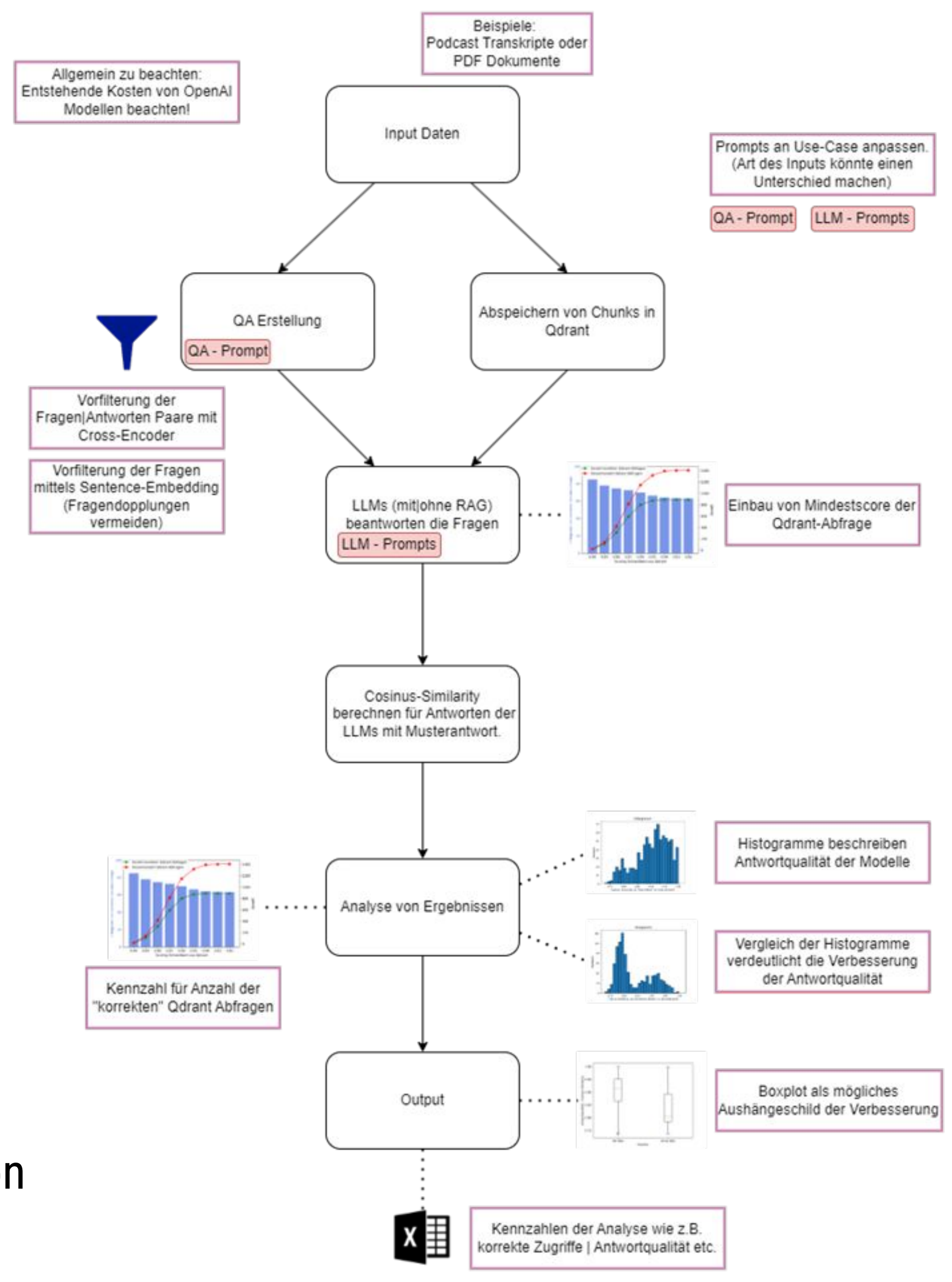
Main obstacles:

- ❑ Laws, especially GDPR (General Data Protection Regulation)
- ❑ Commercialization of Social Media APIs
- ❑ Retrieve data from multiple Social Media Platforms and store/use them in a unified way

QA Process: Question-Answer Pairs



QA-Process: Based on generated Question-Answer-Pairs we evaluate the quality of answers given from the ThWIC Sonar AI. Access to publications yields an improvement on answer qualities.



Beispiele: Podcast Transkripte oder PDF Dokumente
Allgemein zu beachten: Entstehende Kosten von OpenAI Modellen beachten!
Input Daten
Prompts an Use-Case anpassen. (Art des Inputs könnte einen Unterschied machen)
QA - Prompt
LLM - Prompts
QA Erstellung
QA - Prompt
Abspeichern von Chunks in Qdrant
Vorfilterung der Fragen|Antworten Paare mit Cross-Encoder
Vorfilterung der Fragen mittels Sentence-Embedding (Fragendopplungen vermeiden)
LLMs (mit|ohne RAG) beantworten die Fragen
LLM - Prompts
Einbau von Mindestscore der Qdrant-Abfrage
Cosinus-Similarity berechnen für Antworten der LLMs mit Musterantwort.
Analyse von Ergebnissen
Histogramme beschreiben Antwortqualität der Modelle
Vergleich der Histogramme verdeutlicht die Verbesserung der Antwortqualität
Kennzahl für Anzahl der "korrekten" Qdrant Abfragen
Output
Boxplot als mögliches Aushängeschild der Verbesserung
Kennzahlen der Analyse wie z.B. korrekte Zugriffe | Antwortqualität etc.

ThWIC Retreat am 7. und 8. Oktober 2024

Divyasha Sunil Naik, Kevin Lamkiewicz -- FSU Jena / epicinsights (c/o SMA Development GmbH)

CLUSTERS 4 FUTURE
Innovationsnetzwerke für unsere Zukunft

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung