# DaBing

## Application of LLM’s for semantically enriched search & …. Classification (WP10)

Jeldrik Bakker, Olav ten Bosch, Arnout van Delden

10, 11 September, 2025, Warsaw

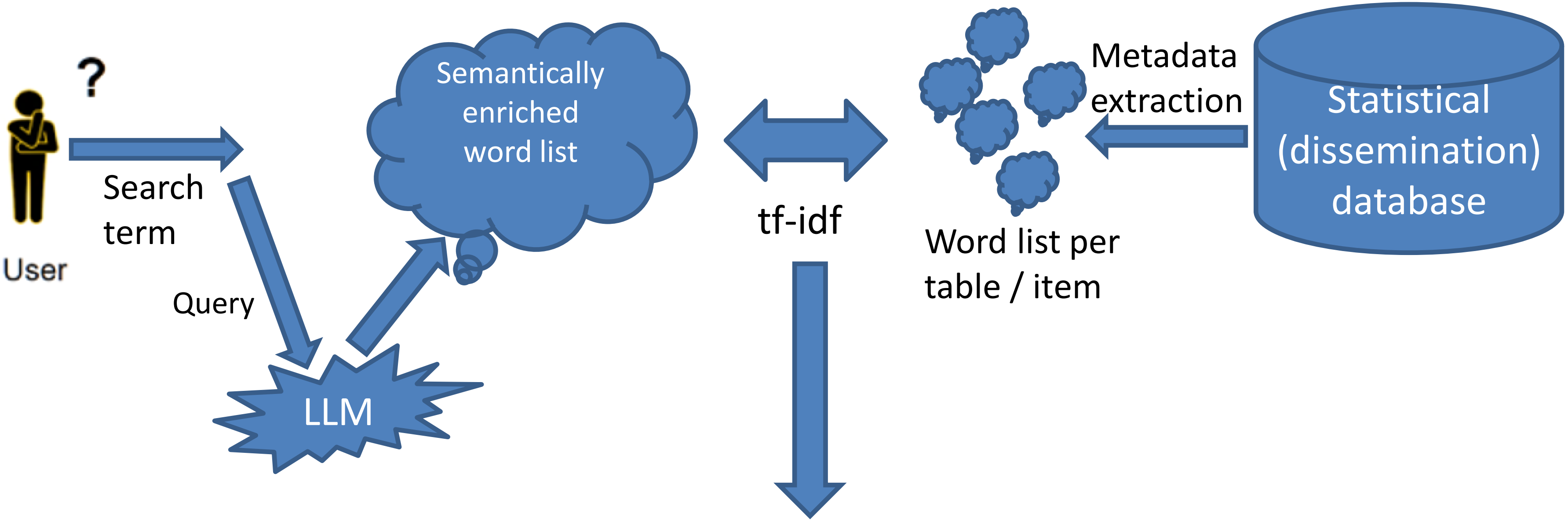
ESSnet AIML4OS, WP10, Classification

# Misclassifications in Statistical Registers: Literature study

- **Problem:**
  - Misclassifications in population frames (e.g., age, economic activity) lead to **biased domain estimates** in official statistics.
  - Common causes: registration errors, time delays, and **definition mismatches**.
- **Literature Approaches:**
  - **Robust machine learning** (Biggio et al., 2011)
  - **Filtering mislabeled instances** (Brodley & Friedl, 1999)
  - **Label noise models** (Eskin, 2002; Rantalainen & Holmes, 2011)
- **Proposed Method:**
  - Combines **machine learning** (for classification) and a **label noise model** (for misclassification detection).
  - Uses a **mixture model** with a **latent variable** to distinguish correct from incorrect labels.
  - Employs a **Generalized EM (GEM)** algorithm for parameter estimation.
- **Application & Validation:**
  - Tested on **NACE codes** in a **statistical business register (SBR)**.
  - Validated using **simulated and real data**.
  - **Open-source Python code** provided (including a case study on dry beans: Koklu & Ozkan, 2020).

Paper in preparation: “Identifying misclassified labels in register data”, A. van Delden, et. al.

# Original concept Dabing (demo in WP12 sep 2024)



?

User

Search term

Query

LLM

Semantically enriched word list

tf-idf

Word list per table / item

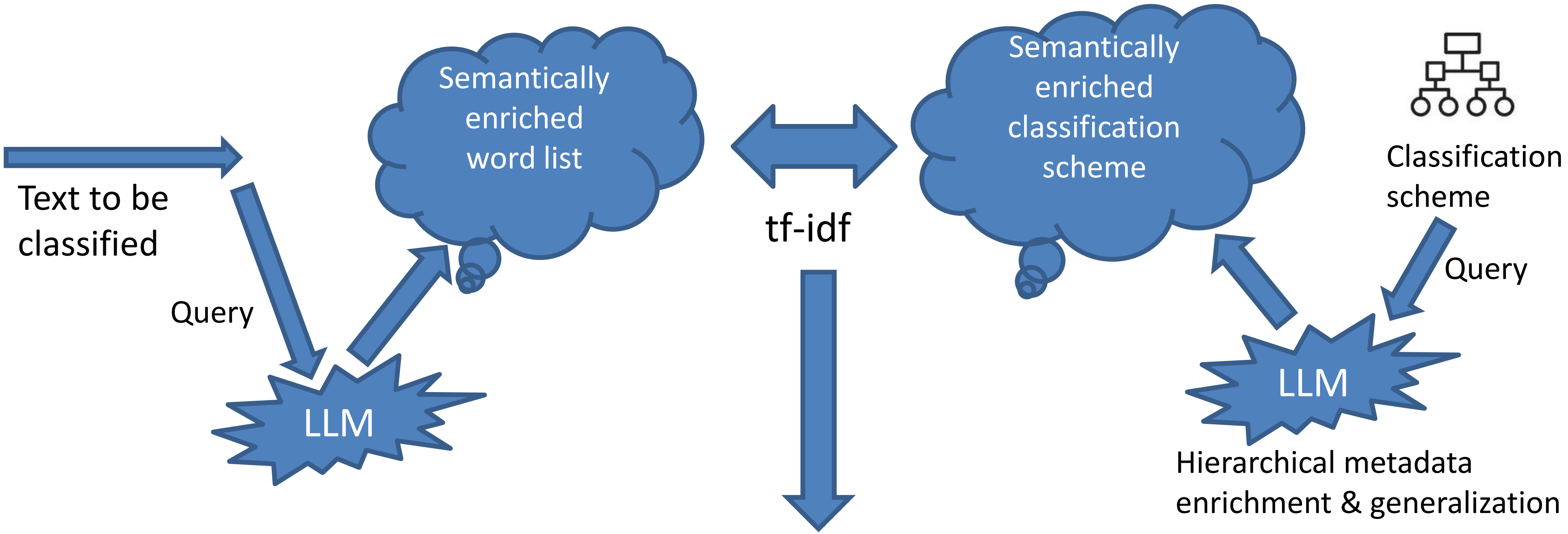
Metadata extraction

Statistical (dissemination) database

| item | score | Link to item |
|---|---|---|
| item1 | 0.67 | link |
| item2 | 0.53 | link |
| item3 | 0.33 | link |

Threshold

# Dabing for coding / classification


Text to be classified
Query
LLM
Semantically enriched word list
tf-idf
Semantically enriched classification scheme
Classification scheme
Query
LLM
Hierarchical metadata enrichment & generalization

| item | score | Link to item |
|---|---|---|
| item1 | 0.67 | link |
| item2 | 0.53 | link |
| item3 | 0.33 | link |

Threshold

# How could this work?

Demo!