

Efficient High-Level Semantic Enrichment of Undocumented Enterprise Data

Markus Schröder

Abstract

In absence of a data management strategy, undocumented enterprise data piles up and becomes increasingly difficult for companies to use to its full potential. As a solution, we propose the enrichment of such data with meaning, or more precisely, the interlinking of data content with high-level semantic concepts. In contrast to low-level data lifting and mid-level information extraction, we would like to reach a high level of knowledge conceptualization. Currently, this can only be achieved if human experts are integrated into the enrichment process. Since human expertise is costly and limited, our methodology is designed to be as efficient as possible. That includes quantifying enrichment levels as well as assessing efficiency of gathering and exploiting user feedback. This paper proposes research on how semantic enrichment of undocumented enterprise data with humans in the loop can be conducted. We already got promising preliminary results from several projects in which we enriched various enterprise data.

Introduction

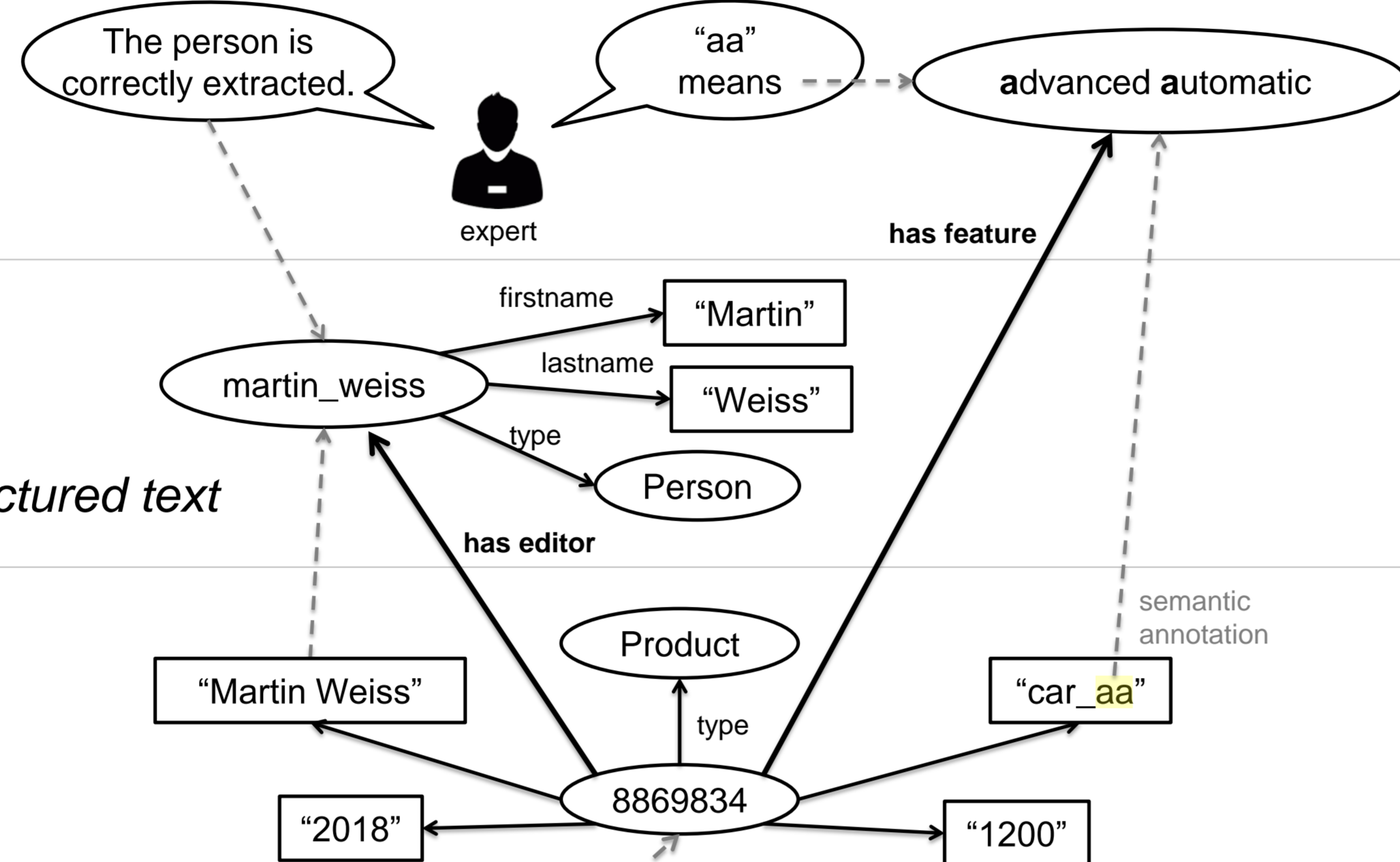
- challenge: messy dataspace
 - increasingly difficult to discover and make use of the data
 - obstacle in performing complex data mining analyses
 - hinders employees to efficiently work with the data content
- approach: semantic enrichment
 - augment data with high-level concepts
 - include feedback from experts via human-in-the-loop approach



High-Level Enrichment: Human-in-the-Loop
Ask domain experts to get user feedback

Mid-Level Enrichment: Information Extraction
Extract structured information from unstructured text

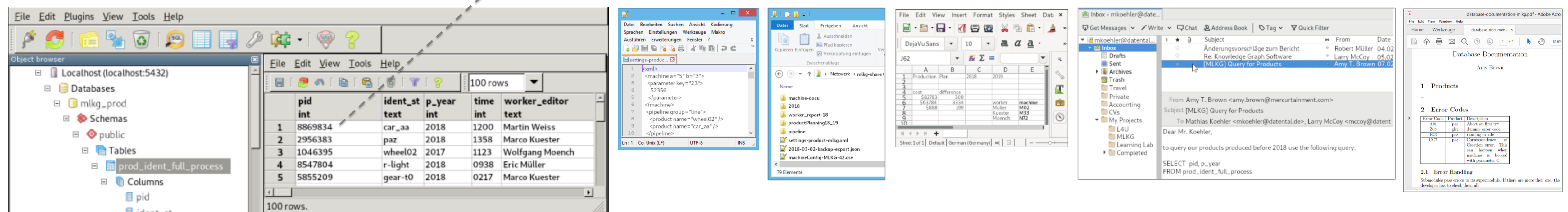
Low-Level Enrichment: Data Lifting
Transform data to RDF using mappings



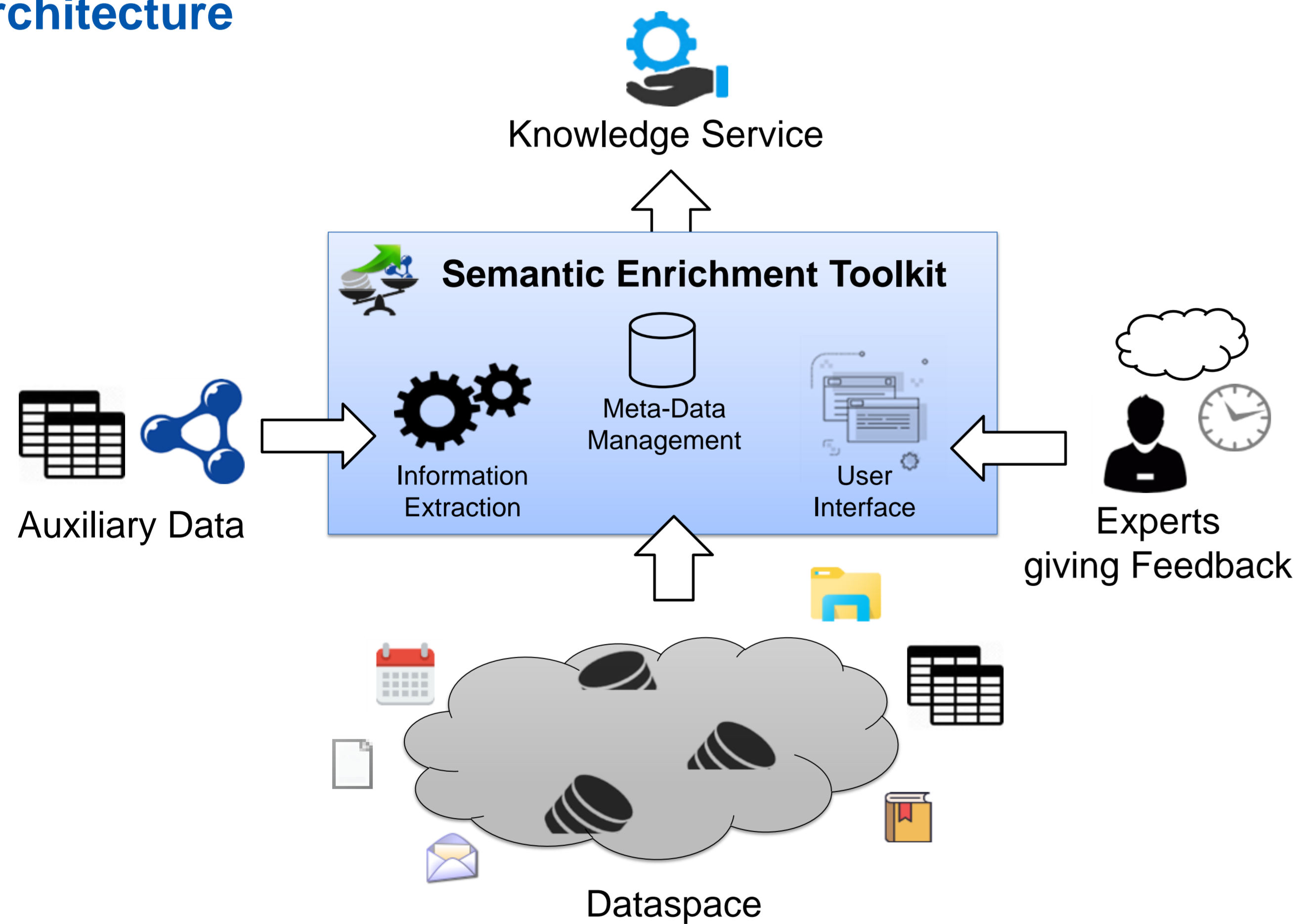
Semantic Enrichment

- semantic enrichment precisely annotates data with meaningful concepts obtained by experts
 - information extraction is not solely data-driven
 - the experts' conceptualisation is not modelled in isolation
- level of enrichment approximates how comprehensive gained knowledge about enterprise data is

Status Quo: Dataspace
Arbitrarily structured data



Architecture



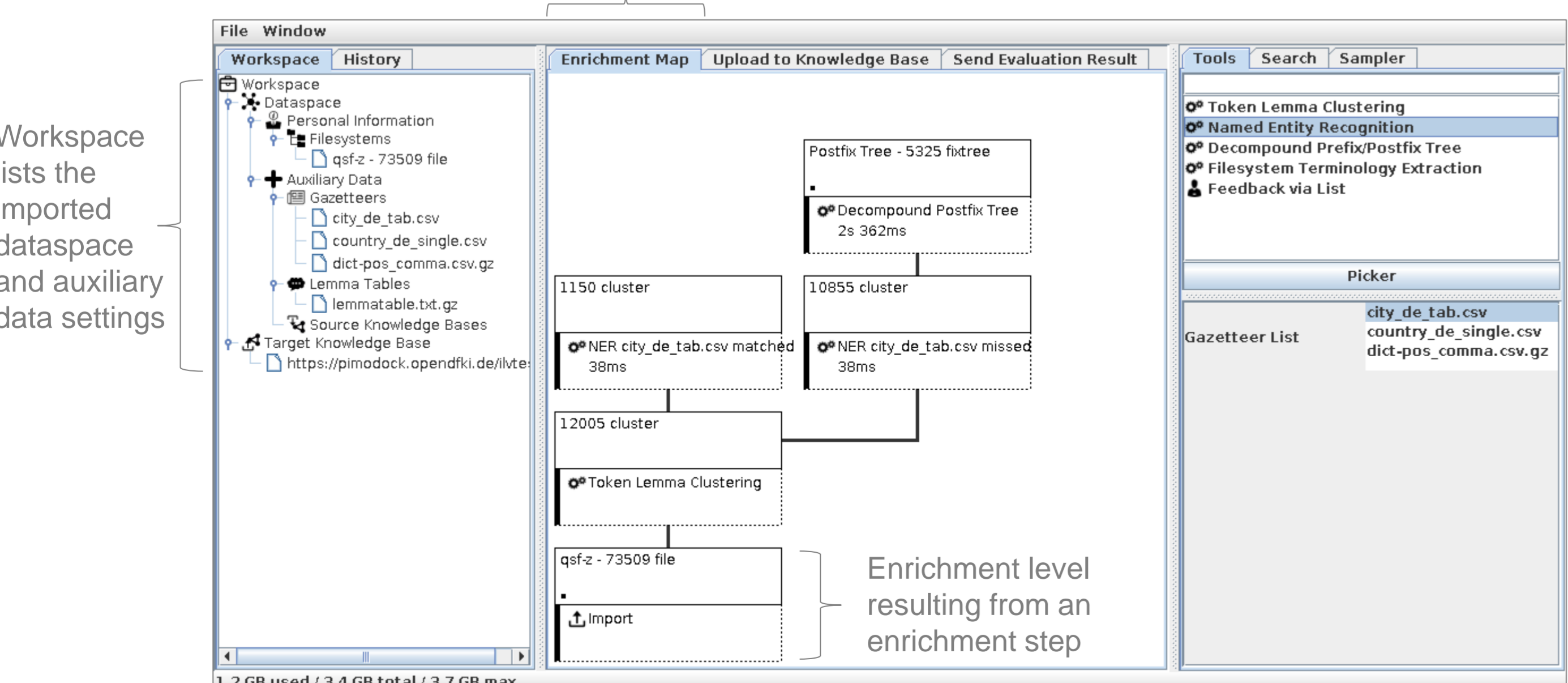
Research Questions

How can an efficient high-level semantic enrichment be conducted?

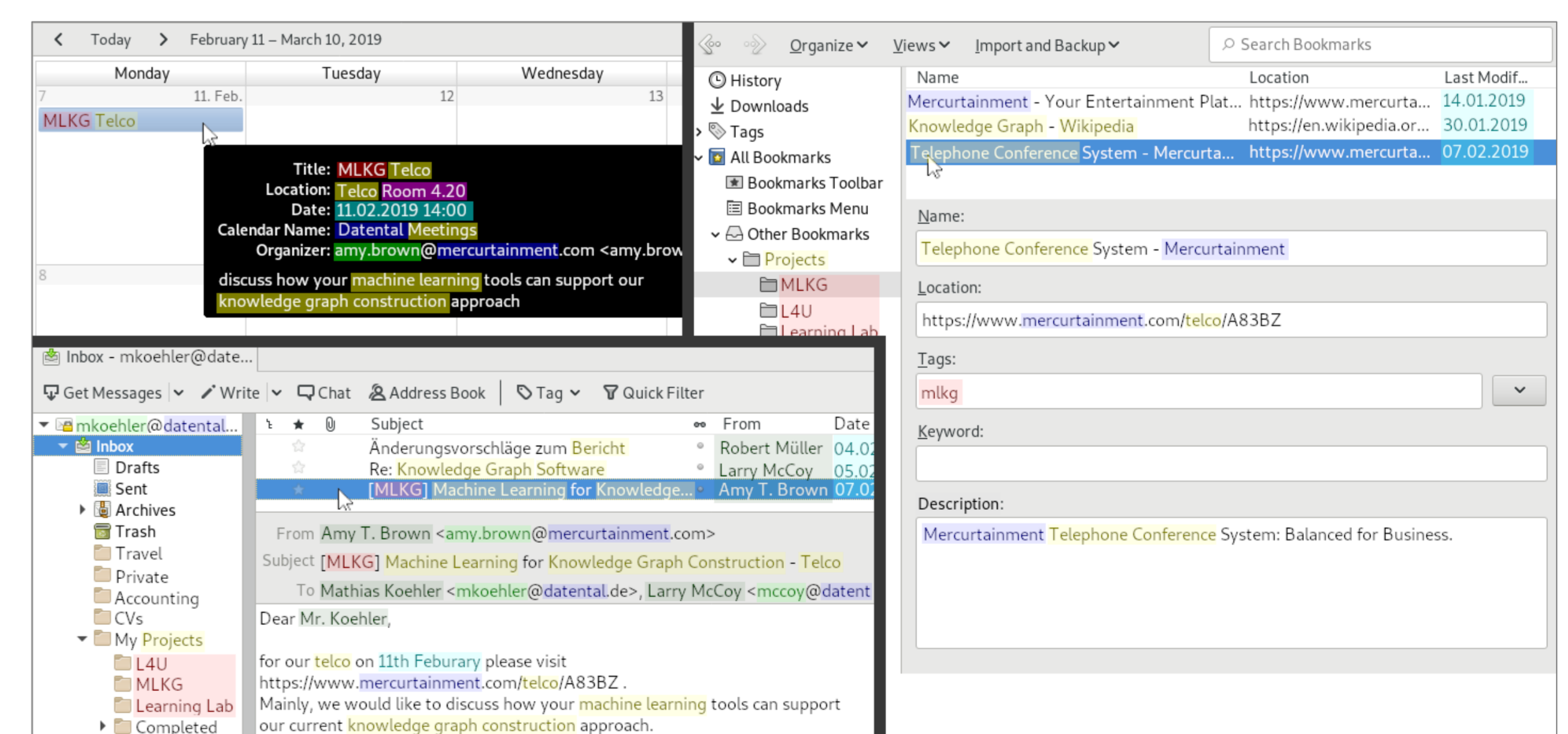
- RQ1: What **state-of-the-art** approaches can be **utilized** and how can they be **adapted**?
- RQ2: How can we **efficiently integrate human experts** in the process to achieve our envisioned high level of enrichment?
- RQ3: How can we **quantify the enrichment level** of state-of-the-art algorithms?
- RQ4: What are appropriate measures to **assess the efficiency** of gathering and exploiting **user feedback**?

Early Prototype: Semantic Enrichment Toolkit

Enrichment Map shows a directed acyclic graph of enrichment levels



Use Case: Interactive Concept Mining on Personal Data



Acknowledgement

Parts of this work have been funded by the German Federal Ministry of Economic Affairs and Energy in the project PRO-OPT (01MD15004D) and by the German Federal Ministry of Food and Agriculture in the project SDSD (2815708615). I thank my doctoral supervisor Prof. Dr. Andreas Dengel and my colleagues Christian Jilek, Dr. Heiko Maus, Dr. Sven Schwarz, Dr. Jörn Hees and Dr. Ansgar Bernardi for their helpful discussions, comments and feedback.



Contact:
Msc. Markus Schröder
Researcher – Smart Data & Knowledge Services
German Research Center for Artificial Intelligence
DFKI GmbH
Phone: +49 631 20575-2070
Mail: markus.schroeder@dfki.de
Website: <http://www.dfki.uni-kl.de/~mschroeder/>