

## *PPP - Text Mining for Metadata Extraction and Semantic Retrieval*

Funded by the [German Academic Exchange Service \(DAAD\)](#)

In cooperation with the [Technical University of Kosice \(TUK\)](#)

Nowadays efficient access to relevant information has become a major challenge in all areas of life. This is especially true for an organizational environment where knowledge can lead to substantial competitive advantage. However, with the quantity of available information growing by the minute, finding out what is actually relevant for a certain information need becomes very tough, especially as vast amounts of information exist in unstructured text documents (e.g. emails, Word or PDF formats) rather than structured databases.

The project aims at developing and evaluating new tools and techniques for fuzzy metadata-based searches and automated metadata extraction. In particular the two goals are:

1. Effective support of metadata-based Information Retrieval using the (extended) Vector Space Model (VSM).
2. Extraction of metadata from text collections using Text Mining techniques such as text classification or clustering.

For the metadata management Semantic Web technologies such as the Resource Description Framework (RDF) and the Dublin Core Metadata standard will be applied. The project will elicit the complementary experience of the partner universities in Košice (Text Mining, Java-based system supporting pre-processing, indexing and mining of text document collections) and Regensburg (Information Retrieval using fuzzy queries on RDF metadata).