Abstract:

The thesis explores the application of approximate string matching in scientific publication record linkage process. An introduction to record matching along with five commonly used metrics for string distance (Levenshtein, Jaro, Jaro-Winkler, Cosine distances and Jaccard coefficient) are provided. These metrics are applied on publication metadata from V3S current research information system of the Czech Technical University in Prague. Based on the findings, optimal thresholds in the F_1, F_2 and F_3-measures are determined for each metric.