# KITspotlight: A System for Spotlighting Researchers in the Media

Michael Färber, Benjamin Zagoruiko, and Markus Wambach

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
michael.faerber@kit.edu
{benjamin.zagoruiko,markus.wambach}@student.kit.edu

**Abstract.** Academic institutions, such as universities, heavily rely on public relations and are thus interested in media monitoring. However, tracking mentions of their researchers in news articles presents challenges such as identifying affiliated personnel and their departments, and aggregating the extracted information. In this paper, we introduce KITspotlight, a novel system that automatically identifies researchers of the Karlsruhe Institute of Technology (KIT) in newspaper articles, associates these individuals with their departments, and presents this information visually. KITspotlight is tailored for both department heads, administrative staff, and individual researchers, focussing on the institution's overall public visibility or individual researcher's public appearance. Analyzing data from 2,280 articles over 12 months, our system offers a model for monitoring academic personnel at any research institution.

**Keywords:** Media Monitoring, Named Entity Recognition, Information Extraction, Text Mining.

## 1 Motivation

Universities and other academic research institutions face significant challenges in public relations, particularly in tracking mentions of their researchers in news articles. This task has traditionally been complex, typically involving manual processes confined to specific departments within institutions. The manual effort to track relevant news articles demands extensive time and resources. In addition, it risks missing key mentions and delayed data compilation, undermining the completeness of the information. In the context of automated information extraction from journalistic texts, notable progress has been made in leveraging named entity recognition (NER) models to identify individuals [1]. However, a critical gap remains: the absence of a comprehensive system that not only accurately associates identified individuals of specific institutions but also effectively filters and aggregates relevant data and presents it visually.

Recognizing these challenges, we have developed a novel system KITspotlight – available online at `http://kit-spotlight.de`[1] – for spotlighting

[1] The code can be found at `https://github.com/michaelfaerber/KITspotlight`

researchers of research institutions in the media. It automates the information extraction from newspaper articles, identifies individuals associated with specific organizations, associates these individuals with their departments (at KIT: institutes), and presents this information in a coherent visual format.

Our system is tailored to address two key user groups: (1) heads of institutions and administration focussing on public outreach, and (2) individual researchers. For the first user group, our system provides a detailed analysis of media presence, crucial for evaluating and refining communication strategies. This feature is essential for managing public perception and orchestrating public relations efforts effectively (see strategic controlling). Meanwhile, individual researchers benefit from a user-friendly tool that tracks their media appearances, offering them concise summaries to actively manage and enhance their external visibility, without the need to read manually news articles for their mentioning. Additionally, the system can be used for providing additional impact metrics for researchers beyond traditional metrics such as the citation count of publications.

## 2 System Design

KITSPOTLIGHT is composed of several components: (1) At its core is a database, created using sqlite3, which functions as the central repository for all data. (2) The Text Analyzing Pipeline processes incoming news articles, performs various analyses, and subsequently stores the resulting data in the database. (3) The frontend, outlined in Sec. 3, is developed with *React* (version 18) and interfaces with the database via a Django API.

Our Text Analyzing Pipeline includes the following data and stages:

**Input Data:** Our input data consists of PDF files that contain multiple articles in a newspaper-style layout. All articles have been pre-filtered to include mentions of the use case institution *Karlsruhe Institute of Technology (KIT)*. For our demo system, we analyzed 2,280 news articles.

**Step 1: Extracting Text.** The first step in our pipeline is the conversion of these PDF files into plain text format, using the Python Library `pdftotext`. For extracting metadata like titles, publishers, and dates, we employ rule-based detection techniques based on regular expressions.

**Step 2: Extracting Persons.** To identify all named entities in the articles, we use a named entity recognition (NER) implementation provided by `spaCy`. The system first determines the article's language using the `langdetect` Python Library. We use the `de_core_news_lg` model from `spaCy` for German articles and the `en_core_web_trf` model for English.

**Step 3: Analyzing Individuals.** We focus in our analysis on individuals employed by our institution. Thus, the next step is to filter out all other individuals mentioned in articles, for instance authors or interview partners. Furthermore, if an individual is affiliated with our institution, we aim to identify the specific institute they are associated with. We develop three distinct methods to determine whether an individual is employed by the institution of

**Table 1.** Evaluation results.

| | Employee Detection | | | Institute Assignment |
|---|---|---|---|---|
| | Precision | Recall | F1-Score | Accuracy |
| Google-Search | 0.8166 | 0.9650 | 0.8846 | 58% |
| SemOpenAlex | 1.0000 | 0.8218 | 0.9022 | 0% |
| Staff-Directory | 0.9942 | 0.9884 | 0.9913 | 78% |

interest: (1) *Google Search:* This method uses a Google search with the individual's name in quotes and "kit" to filter for KIT-related webpages. The top search results are then analyzed. (2) *SemOpenAlex Search:* This approach queries SemOpenAlex [2], a large, up-to-date academic knowledge graph. It helps identify researchers but not their specific KIT-institute. (3) *Staff-Directory Search:* This approach accesses KIT's staff directory through POST requests, yielding relevant matches in a JSON array. Privacy settings control the visibility of personal details, with certain information like institute affiliation potentially concealed.

**Evaluation.** We performed an evaluation based on 174 names selected randomly from news articles, comprising 50 KIT employees and 124 non-KIT employees. The results are given in Table 1. The Staff-Directory method demonstrates the highest F1-score in identifying KIT employment, but privacy settings occasionally obscure institute details, underscoring the relevance of alternate approaches. Both the Google-Search and SemOpenAlex methods showed strong F1-scores. However, SemOpenAlex often lacks institute information. Despite frequently including researchers' ORCID identifiers, the institute data is not always available or differs from the Staff-Directory format. In conclusion, our system starts with a lookup within the Staff-Directory to verify if an individual is working at KIT. Should the result lack institute details, we proceed with an additional Google query to try sourcing the institute from the URL of the researchers' institutional-homepage.

With around 1,600 articles analyzed, we successfully identified 472 mentioned individuals working at KIT, which in total were mentioned 1,979 times. We also achieved to match 78% of the individuals to an institute.

## 3 User Interaction

**Front Page.** Fig. 1 showcases the main page. A bar chart positioned in the top left tracks the frequency of mentions of KIT employees over time. Adjacent to this, on the right, another chart depicts the proportional occurrence of KIT institutes. In the bottom left, a table lists recent articles with their titles and the KIT employees mentioned therein. In the bottom right corner, a pie chart shows how often different media outlets mention the identified researchers.

**Institute-Page.** Information of media coverage for specific KIT institutes is given at the respective pages (see Fig. 2), showing, among other things, a time series analysis of the individuals belonging to the institute.

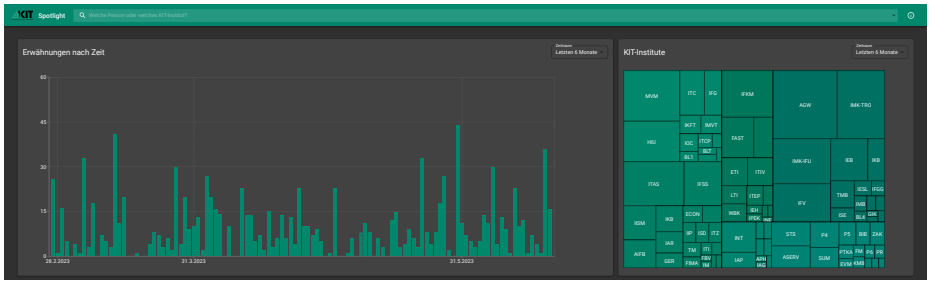


**Fig. 1.** Part of KIT Spotlight's welcome page with overall statistics

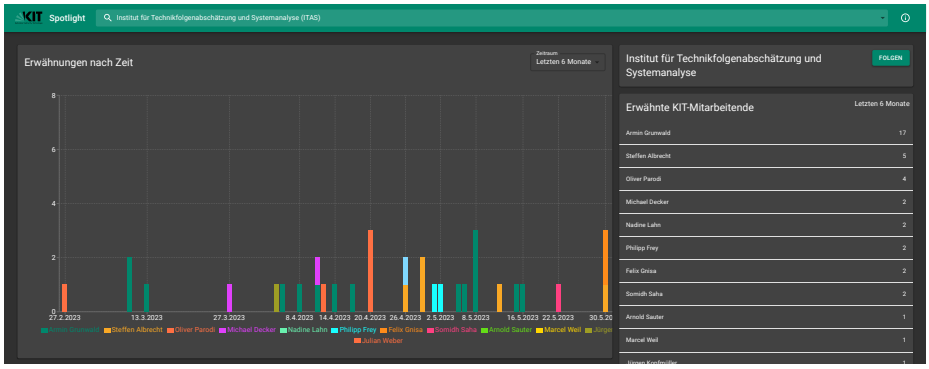


**Fig. 2.** Page of the KIT Institute for Technology Assessment and Systems Analysis

**Person-Page.** On the individual's profile page, we include a comprehensive listing of articles in which the person has been mentioned, showing the article titles, the publisher names, and the URLs.

**Follow-Function.** In addition to viewing and exploring the data on the website, the user has the option to subscribe to an institute or person and get notified via email when a news article mentions the followed entity.

## 4 Conclusion

In this paper, we developed a system that identifies researchers in news articles, simplifies press monitoring in academia, and serves as a blueprint for global academic media monitoring.

## References

1. Buz, C.: Validierung eines NER-Verfahrens zur automatisierten Identifikation von Akteuren in journalistischen Texten (2021). https://doi.org/10.5445/IR/1000131532
2. Färber, M., Lamprecht, D., Krause, J., Aung, L., Haase, P.: SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples. In: Proc. of ISWC'23 (2023)