new/s/leak

network of searchable leaks

Information Extraction and Visualization for Investigative Data Journalists

Seid Muhie Yimam † Marcel Rosenbach ‡ Michaela Regneri ‡ Franziska Lehmann ◊ Chris Biemann †
Heiner Ulrich ‡ Tatiana von Landesberger ◊ Alexander Panchenko † Uli Fahrer † Kathrin Ballweg ◊

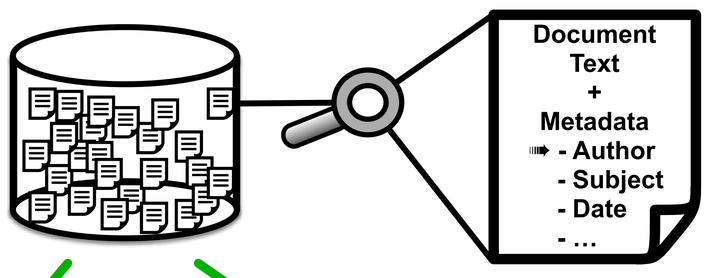
† FG Language Technology, Dept. of Computer Science, Technische Universität Darmstadt

‡ SPIEGEL-Verlag Hamburg, Germany

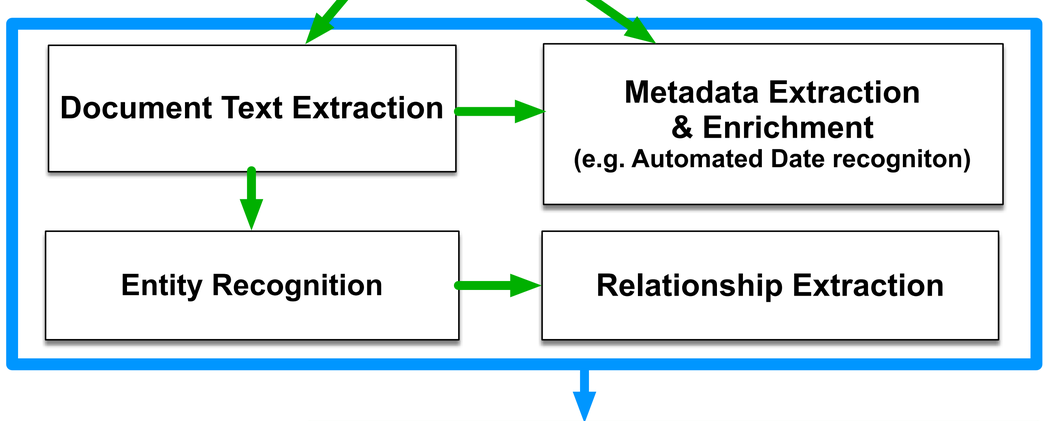
◊ Graphic Interactive Systems Group, Dept. of Computer Science, Technische Universität Darmstadt

Input: Document Collection

messy, unstructured data

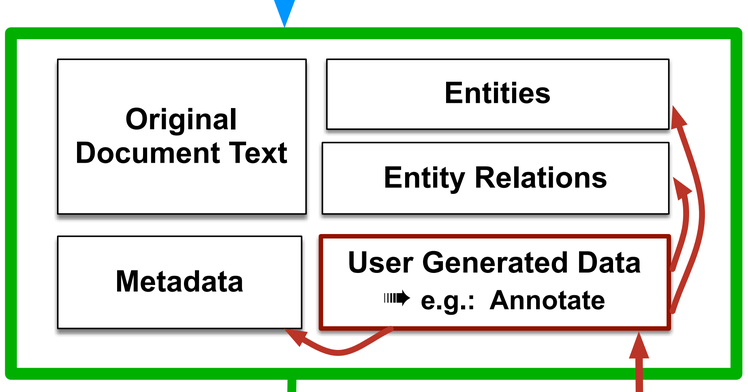


Backend



NLP Output = Visualization Input

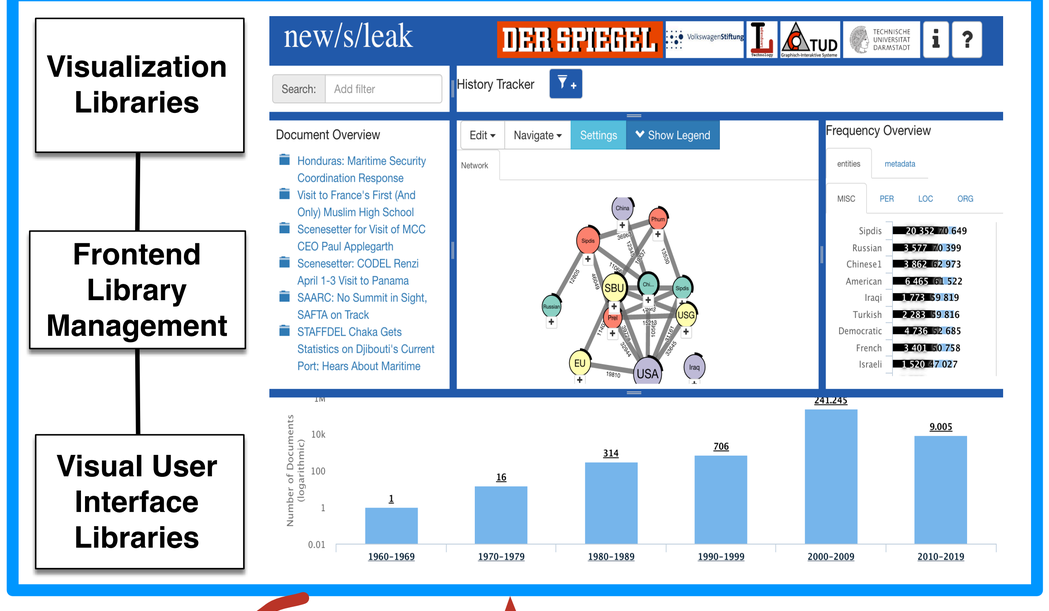
preprocessed, structured data



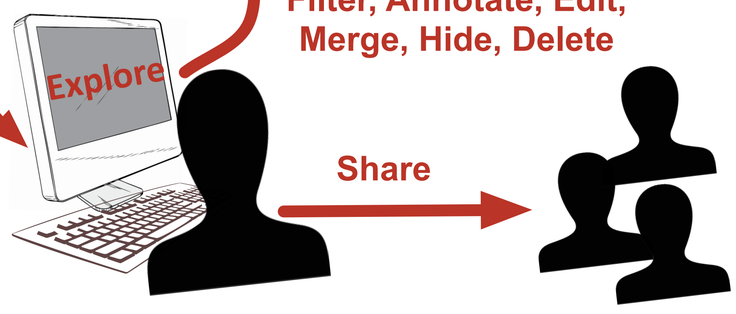
REST-API

Annotations, new Entities

Frontend



User Interaction



Filter, Annotate, Edit, Merge, Hide, Delete

Share

new/s/leak – Overview

- Supports investigative journalism
- Designed for large textual datasets (un- or semi-structured)
- Generic infrastructure, easy to adapt to new datasets

Key Requirements for Journalists

- Identify key entities
- Guided document browsing
- Analyze entity connections
- Temporal development of documents
- Geographic distribution of events
- Annotate documents
- Save and share findings

Backend

- Pre-process and clean documents
- Extract generic and dynamic metadata using latest NLP technologies
- Extract relationships between entities
- All extractions are done with the Apache Spark cluster computing framework for parallel computations
- Store data in PostgreSQL database and Elasticsearch indexes
- Provide API services for queries and updates

Frontend

- Support interactive visualization with different views
- **Graph view**: shows named entities and their relations, annotate and update the graph
- **History**: tracing / reverting user interactions
- **Timeline**: document evolution over time
- **Document view**: list, read & annotate documents
- **Full text search**
- **Metadata filter**: constrained data exploration

Project Blog: http://newsleak.io/
Demo: http://bev.lt.informatik.tu-darmstadt.de/newsleak/
Source code: https://github.com/tudarmstadt-lt/newsleak/

Seid Muhie Yimam, Heiner Ulrich, Tatiana von Landesberger, Marcel Rosenbach, Michaela Regneri, Alexander Panchenko, Franziska Lehmann, Uli Fahrer, Chris Biemann and Kathrin Ballweg (2016): new/s/leak – Information Extraction and Visualization for an Investigative Data Journalists. ACL 2016 Demo Session, Berlin, Germany

The authors are grateful to data journalists at Spiegel Verlag for their helpful insights into journalistic work and for the identification of tool requirements. The authors wish to thank Lukas Raymann, Patrick Mell, Bettina Johanna Ballin, Nils Christopher Boeschen, Patrick Wilhelmi-Dworski and Florian Zouhar for their help with system implementation and conduction of the user study. The work is being funded by Volkswagen Foundation under Grant Nr. 90 847.