Jens Aurich (PhD Researcher, jens.aurich@iisg.nl)
International Institute of Social History (Amsterdam)

Collecting Strike Data from Digitized Historical Newspapers

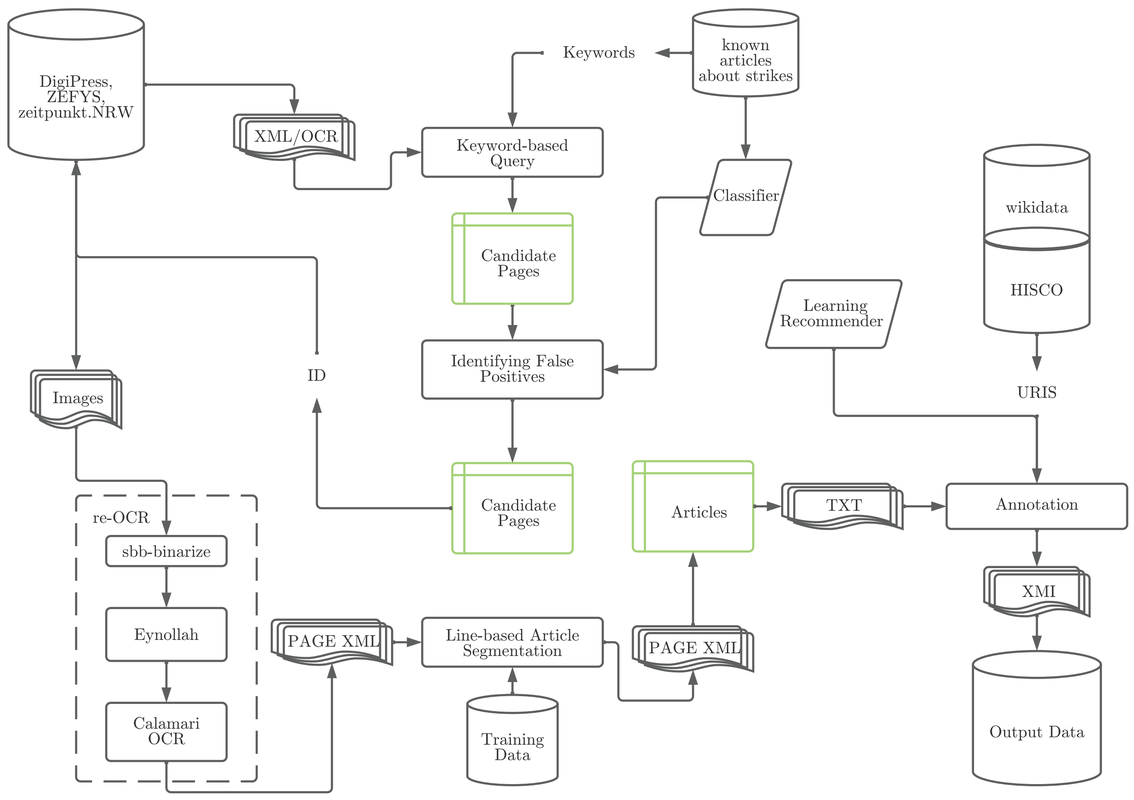
Potential and Challenges of a Digital Workflow

I create and implement a workflow for the retrieval and annotation of text relevant for the study of strikes from three major collections of digitized German historical newspapers.

To the growing Digital History literature concerned with the exploitation of large collections of digitized newspapers, this project contributes a proof of concept for the suggested workflow, an in-depth quality review of a selection of newspapers from three major German digital collections, a re-OCR pipeline and a segmentation approach to separate news reports.

To the historiography of strikes in the first two decades of the German Empire, it contributes a novel data resource and a first assessment of the coverage of domestic labor conflicts in major newspapers.

Workflow



Why Do We Care?

Global History of Collective Labor Action and Digital History

The global turn in labor history led to a conceptual, temporal and geographic expansion of the historiography. New histories of labor and exploitation go beyond the conventional locus of European industrial capitalism of the 19th century. This shift has important implications for the study of workers' resistance against exploitation – collective labor action. Against labor history's conventional concern with strikes by industrial workers, a new range of actions and actors are now brought into focus: Boycotts, sabotage, desertions, riots or the creation of mutual funds by indentured, enslaved, unemployed, self-employed or domestic workers are examples.

For this research agenda, Global Labor History requires new tools and workflows to collect data from a vast source base beyond the classical strike statistics collected by states and unions. This project tests the potential of collecting data on collective labor action from digitized daily newspapers using the toolkit of Digital History.

Results

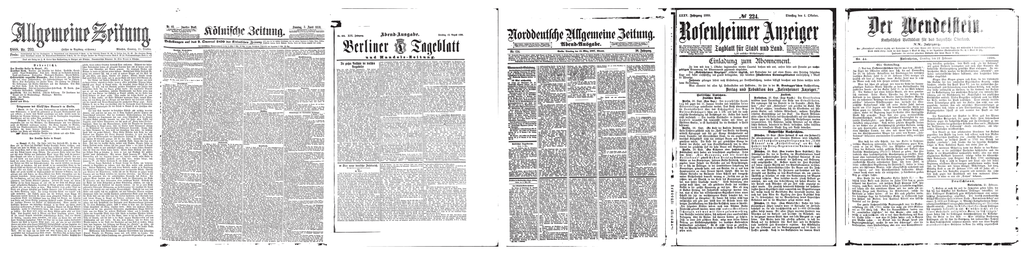
- Proof of concept workflow
- Corpus of 21,000 candidate articles, 4,000 annotated with metadata
- Reporting on strikes has international scope but is biased towards large strikes in central industries
- The selected newspapers report only 10% of strikes recorded in existing microdata for 1871-1875

Method

0 Data

digiPress (digipress.digitale-sammlungen.de) · zeit.punkt NRW (zeitpunkt.nrw) · ZEFYS (zefys.staatsbibliothek-berlin.de)

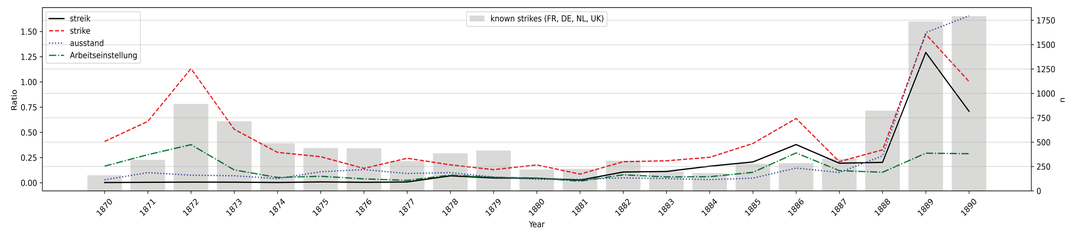
The data consists of 42,000 daily issues (1870-1890) of six newspapers selected from the three largest collections of retrodigitized historical German newspapers for the late nineteenth century. The selection of newspaper titles follows existing newspaper-based studies of social unrest in the 19th Century and includes two major interregional newspapers and four regional newspapers published in Berlin and Bavaria.



1 Corpus Creation

Word Embeddings (impresso-project.ch) · fastText (fasttext.cc)

I create a subcorpus of candidate newspaper pages by identifying relevant historical keywords in a bottom-up manner using a statistical approach (TF/IDF) and a novel ground truth dataset of newspaper articles collected based on source references in existing quantitative sources. To increase the accuracy of the corpus, I identify false positives with a text classifier trained on a tagged set of 2000 random sentences embedded using a FastText model with subword features (F-score = 0.91).

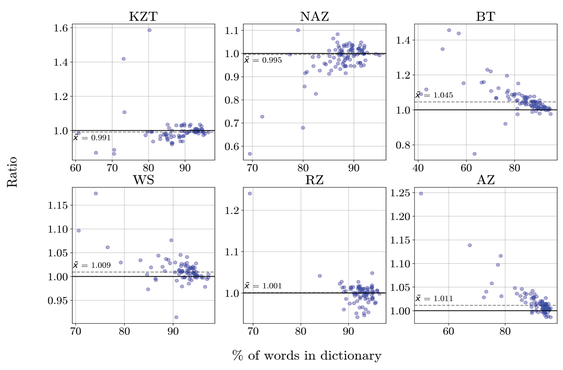


The ratio of keywords per issue and the number of known strikes by year

2 Improving OCR & OLR

sbb-binarize (github.com/qurator-spk/sbb_binarization) · Eynollah (github.com/qurator-spk/eynollah) · Calamari OCR (github.com/Calamari-OCR/calamari) · Image Annotator (VIA) (robots.ox.ac.uk/~vgg/software/via/)

I implement a re-OCR pipeline using pre-trained models that improves the mean OCR quality of the entire corpus by 1.2%. Unlike this moderate improvement in character recognition, the layout and reading order recognition is substantially improved by the re-OCR pipeline, reducing the share of pages with omitted lines (1% compared to 15% in the provided OLR) or an incorrect reading order (14.6% instead of 58%).
To meet the challenge of missing article segmentation, I implement a line-based classifier trained on the visual and textual features of an annotated set of 256 pages that performs well across the selected newspapers (F-score: 0.97).



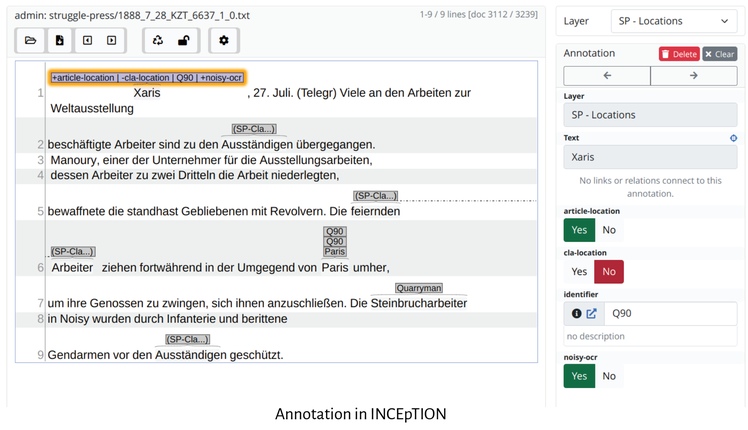
OCR Quality Review

Creating Line Segmentation

3 Annotation

INCEpTION (github.com/inception-project/inception) · HISCO (historyofwork.iisg.nl) · wikidata (wikidata.org)

I annotate a sample of the resulting subcorpus of 21,000 relevant newspaper articles using INCEpTION.
On the document level, I annotate whether or not a given article is a false positive and whether it mentions a past or ongoing event of collective labor action. Relying on INCEpTION's built-in learning recommender systems and SPARQL query capabilities, I annotate places and occupations mentioned in the context of a conflict event and link them to their corresponding URIs.

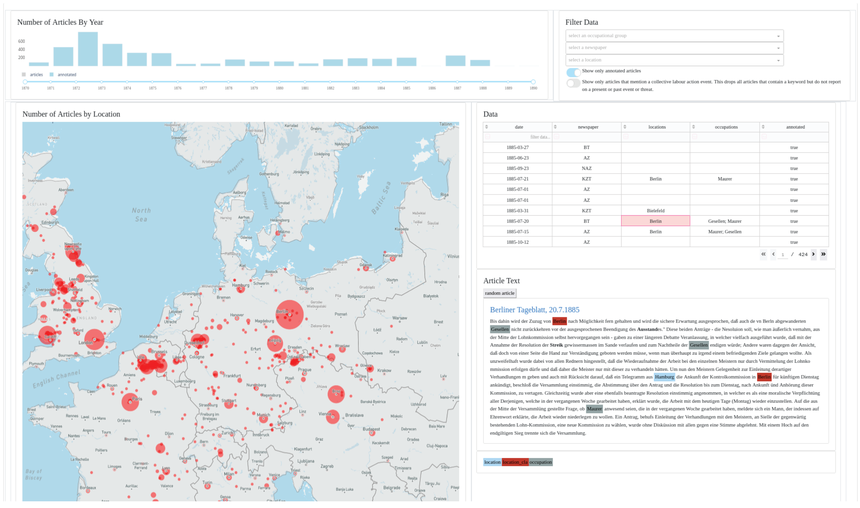


Annotation in INCEpTION

4 Visualization

Plotly Dash (plotly.com/dash/)

An interactive visualization of the resulting data allows for basic exploration of the corpus and features filtering by occupations, locations or newspapers. Articles are displayed with rendered annotations and reference to the original scan image provided by the archives.



HUMBOLDT-UNIVERSITÄT ZU BERLIN · international institute of social history