

**Staatsbibliothek  
zu Berlin**  
Preußischer Kulturbesitz



# Newspapers as Data

What's the News for AI and DH?

Clemens Neudecker | Berlin State Library  
ÖNB Labs Symposium 2024 | 25 November 2024

10 years ago...

## Europeana Newspapers Information Day @ ÖNB (16 October 2014)



<http://www.europeana-newspapers.eu/information-day-at-the-austrian-national-library/>

# Europeana Newspapers

- Europeana Newspaper gathered 20+ million newspaper pages in a single access point, 12 million pages processed with OCR



The image is a project fact sheet for Europeana Newspapers. It features a large, stylized 'e' logo on the left, composed of various newspaper clippings. The title 'Europeana Newspapers' is prominently displayed in a large, black, serif font at the top center. Below the title, the text 'Special Issue - Final Report' is written in a smaller, black, serif font. The main body of the fact sheet is a collage of various historical newspaper pages, showing different layouts, headlines, and text. On the right side, there is a logo for 'ICTPSP' and the European Union flag. The date '20 August 2015' is printed below the flag. The text 'Free and Open Source Software Tools for Newspapers' is written in a bold, black, sans-serif font. At the bottom, the text '20+ million newspaper pages available online' is written in a bold, black, sans-serif font. The overall design is clean and professional, with a focus on the historical and digital aspects of the project.

**Europeana Newspapers**

Special Issue - Final Report

The Europeana Newspapers project ([www.europeana-newspapers.eu](http://www.europeana-newspapers.eu)) is an EU-funded Best Practice Network to make European digital historical newspapers available via Europeana ([www.europeana.eu](http://www.europeana.eu)) & The European Library ([www.theeuropeanlibrary.org](http://www.theeuropeanlibrary.org)).

20 August 2015

**Free and Open Source Software Tools for Newspapers**

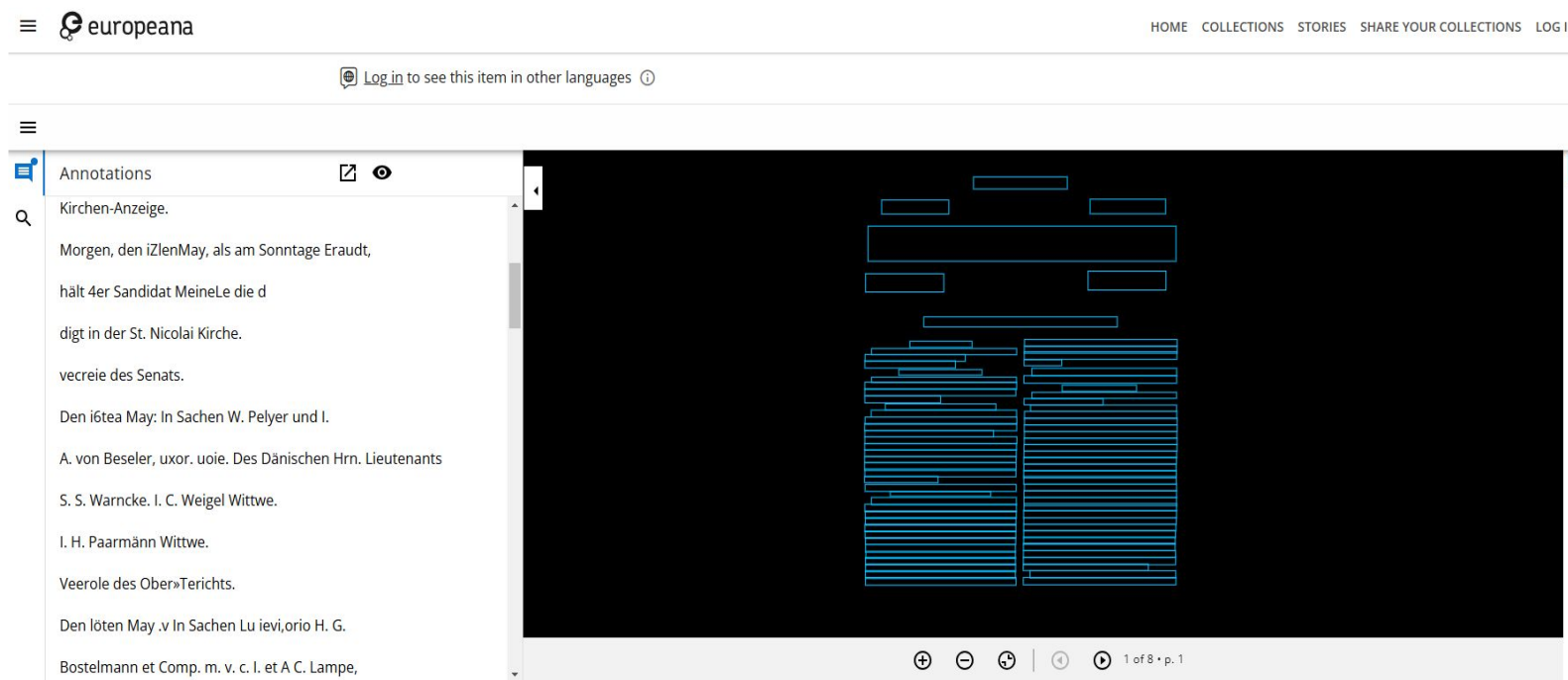
The Europeana Newspapers project has developed a number of free and open source software tools. These tools can help libraries and other institutions to refine their digitised historical newspapers and to evaluate the quality

**20+ million newspaper pages available online**

...meanwhile...

# Europeana Newspapers

- Europeana Newspaper thematic collection still exists, but is in decay, with technical/usability issues and less content than originally published
- The quality of the OCR (produced in 2012-2014) is sub-optimal...





# Deutsches Zeitungsportal

- German “copy” of Europeana Newspapers, launched 10/2021
- Strong growth of content (~26m pages), but few OCR available

[illegible]

@Stabi Berlin...



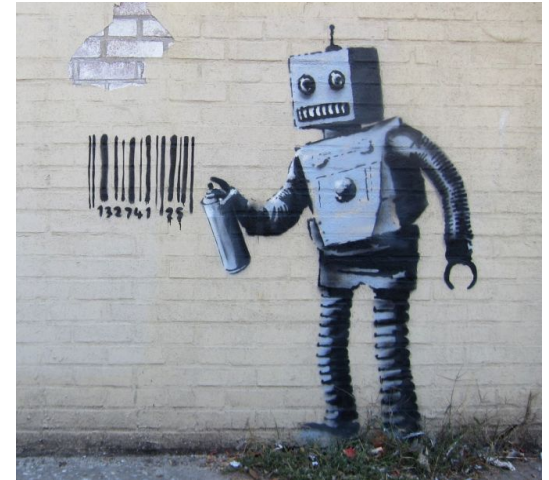
# Staatsbibliothek zu Berlin - Berlin State Library (Stabi)

- is part of the Prussian Cultural Heritage Foundation - Stiftung Preußischer Kulturbesitz (SPK)
- is open freely to the public 7 days/week in two sites in Berlin
- is collecting scientific literature in all languages and from all times and countries since 1661
- has a collection of ca. 12M books with an annual growth of approx. 100k titles
- **Digitized Collections** provide access to >220,000 digitized documents under Public Domain license and via IIF API
- Continuous newspaper digitisation with OCR in **Zefys**
- **Stabi Lab** for experiments, events, datasets, digital humanities



# AI (or rather Machine Learning) at Stabi Berlin

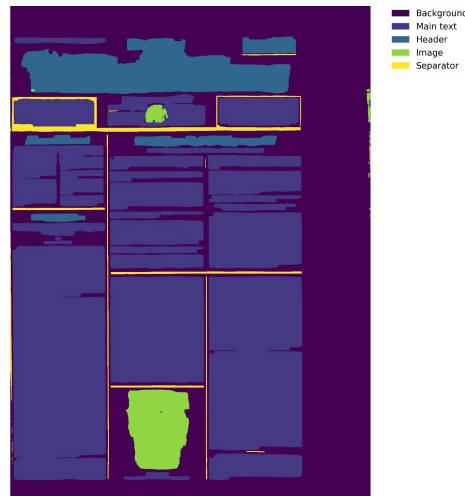
- Driven by third-party funded projects
  - 2016—2024: **OCR-D** (DFG)
  - 2018—2021: **Qurator** (BMBF)
  - 2022—2025: **Mensch.Maschine.Kultur** (BKM)
- Team
  - 7x 100% Machine Learning Engineer
  - 1x 50% Library and Information Scientist
  - 1x 100% Research Data Manager / Historian
- Hardware
  - 2x GPU Servers NVIDIA Tesla A100
- Goals
  - Research and development of open source technologies for cultural heritage
  - Search and retrieval across all collections, for text and image content
  - Provision of collections as open and machine-readable digital data (**Collections as Data**)
  - Responsible use of AI and curation of digital data under consideration of problematic content based on ethical, legal and social criteria



CC-BY-SA Scott Lynch via [Flickr](#)

# Document Layout Analysis (Segmentation)

- *Document Layout Analysis with Deep Learning and Heuristics*, 2023.  
<https://doi.org/10.1145/3604951.3605513> | <https://github.com/qurator-spk/eynollah>



# Image Similarity and Text-Image-Search

- *Gauging the Limitations of Natural Language Supervised Text-Image Metrics Learning by Iconclass Visual Concepts*, 2023.

<https://doi.org/10.1145/3604951.3605516> | [https://github.com/qurator-spk/sbb\\_images](https://github.com/qurator-spk/sbb_images)

### Description Search Results

Image

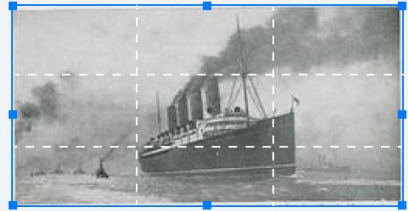
Description

PPN

Big ship entering port

Enter image description. **English only!**

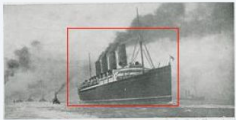
Your search:



Images matching the description you entered:


Search Similar Images

View in Digitized Collections




Search Similar Images

View in Digitized Collections



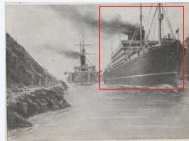
Search Similar Images

View in Digitized Collections




Search Similar Images

View in Digitized Collections




Search Similar Images

View in Digitized Collections



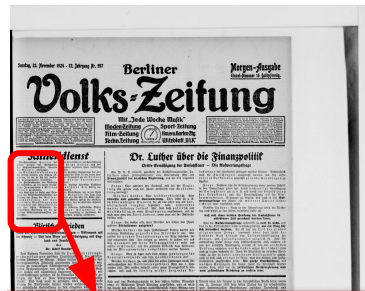
Search Similar Images

View in Digitized Collections



# Text Recognition (OCR, HTR)

- *OCR-D: An end-to-end open source OCR framework for historical printed documents*, 2019.  
<https://doi.org/10.1145/3322905.3322917> | <https://github.com/OCR-D/core>



Ministerpräsident Herriot hat  
gestern morgen den Vorsitzenden  
der Rußland-Kommission, Senator  
de Monzie, empfangen.  
Die Botschafterkonferenz  
hat gestern morgen eine Sitzung ab-  
gehalten, in der sie sich mit laufen-  
den Angelegenheiten beschäftigte.  
Die Handelskammer von Bordeaux  
hat beschlossen, zu der neuen franzö-

Ministerpräsident Herriot hat  
Die Feierlichkeiten zur Ueberrührung  
geiern morgen den Vorsitzenden  
der Autlan-Kommission, Senator  
di Monzie, empfangen.  
Die Botschafterkonferenz  
bat gestern morgen eine Sitzung ab-  
gehalten, in der sie sich mit laufen-  
ten Angelegenheiten beschäftigte.  
Die Handelskammer von Bordeaux  
bat beschloffen, zu der neuen franzö-  
sischen Inlandsanleihe 1 Million  
Francs zu zeichnen.  
Aus Genf sind in Sofia zwei Dele-  
gierte der Balkendundskommisfion zur  
Prüfung der Frage der Massen-  
auswanderung der bul-  
garischen Bevölkerung aus  
Thrazien und Mazedonien  
und der letzten Beschwerde der  
bulgarischen Regierung an die zu-  
ständige Dölkerbundskommisston ein-  
getroffen.

2014

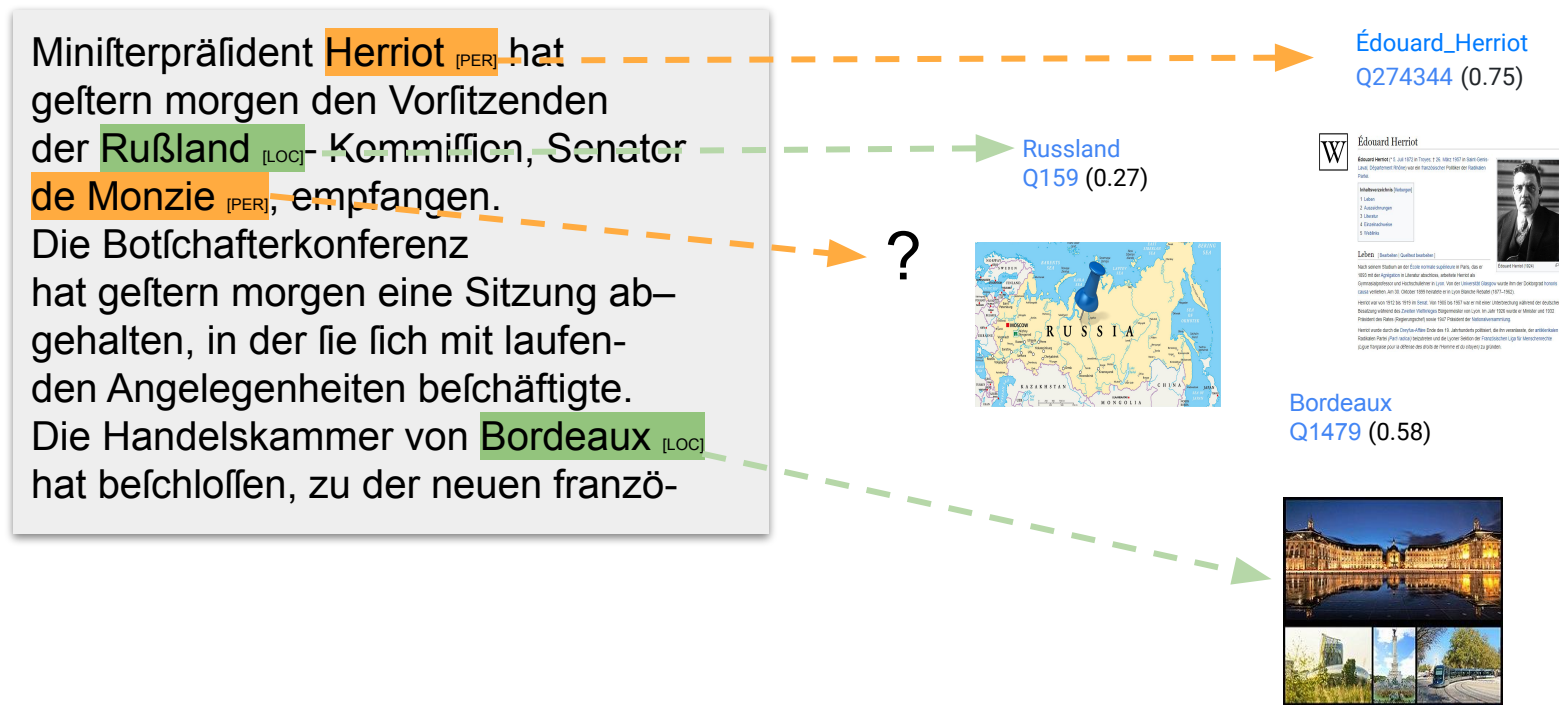
Ministerpräsident Herriot hat  
geiern morgen den Vorsitzenden  
der Rußland- Kommission, Senator  
de Monzie, empfangen.  
Die Botschafterkonferenz  
hat gestern morgen eine Sitzung ab-  
gehalten, in der sie sich mit laufen-  
den Angelegenheiten beschäftigte.  
Die Handelskammer von Bordeaux  
hat beschloffen, zu der neuen franzö-  
sischen Inlandsanleihe 1 Million  
Francs zu zeichnen.  
Aus Genf sind in Sofia zwei Dele-  
gierte der Völkerbundskommisfion zur  
Prüfung der Frage der Massen-  
auswanderung der bul-  
garischen Bevölkerung aus  
Thrazien und Mazedonien  
und der letzten Beschwerde der  
bulgarischen Regierung an die zu-  
ständige Völkerbundskommisfion ein-  
getroffen.

2024



# Named Entity Recognition and Entity Linking

- *BERT for Named Entity Recognition in Contemporary and Historic German*, 2019.  
[https://konvens.org/proceedings/2019/papers/KONVENS2019\\_paper\\_4.pdf](https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf) |  
[https://github.com/qurator-spk/sbb\\_ner](https://github.com/qurator-spk/sbb_ner) | [https://github.com/qurator-spk/sbb\\_ned](https://github.com/qurator-spk/sbb_ned)





...elsewhere...

# Newspaper Navigator

- LC LABS Newspaper Navigator Dataset and Visual Similarity Search



LABS

About **Work** Events Data for Exploration AI at LC

« Experiments

## Newspaper Navigator



Search historic newspaper photos!

Explore the visual and textual content within the Chronicling America digitized newspaper collection in new ways using machine learning.

Dataset and Search Application now live!

### The Newspaper Navigator Search Application

This experimental web application allows you to browse over 1.56 million images extracted from the [Chronicling America](#) database of digitized historic newspapers using machine learning. Now, you can use [this tool](#) to search the images by visual similarity by training your *own* machine learning classifiers!

### The Newspaper Navigator Dataset

The [Newspaper Navigator dataset](#) is the outcome of the first phase of Ben Lee's project. It consists of extracted visual content for 16,358,041 historic newspaper pages in *Chronicling America*.

# Chronicling Germany

- Replicating “Chronicling America Newspaper Navigator” with German data drawn from regional newspaper portal [zeitpunkt.nrw](https://zeitpunkt.nrw)

Name	Letzter Commit	Letzte Aktualisierung
data	update data	vor 2 Monaten
documents	update data	vor 2 Monaten
models	update model	vor 1 Monat
script	clean up	vor 2 Monaten
work_in_progress	clean up	vor 2 Monaten
.gitattributes	attributes	vor 9 Monaten
LICENSE	Update file LICENSE	vor 5 Monaten
readme.md	update readme	vor 3 Wochen

readme.md

## Chronicling Germany historical newspaper dataset

This german historical newspaper dataset contains 581 annotated pages. Layout regions are manually annotated, while transcriptions and line polygons are automatically generated and then corrected. All annotations are provided in the PAGE-XML format, one file for each page. Image and annotation files have the same name. Source of the images is <https://zeitpunkt.nrw/>. The folder data contains the images and annotations folder, which can be used for training. The split.json file provides a dataset split, which associates file stems with Training, Validation and Test datasets. The Test dataset contains out of distribution newspaper pages. To be able to evaluate those specifically, the split.json file provides the list of only these test dataset files within the Out of distribution test list. We provide a models folder and a dataset split.json, to make comparisons with our results possible. Github <https://github.com/Digital-History-Bonn/Chronicling-Germany-Code>

## Chronicling Germany: An Annotated Historical Newspaper Dataset

**Christian Schultze**  
High-Performance Computing  
and Analytics (HPCA-Lab)  
Universität Bonn

**Niklas Kerkfeld**  
HPCA-Lab, Universität Bonn

**Kara Kuebart**  
Institut für Geschichtswissenschaft  
Universität Bonn

**Princilia Weber**  
Institut für Geschichtswissenschaft  
Universität Bonn

**Moritz Wolter\***  
HPCA-Lab, Universität Bonn

**Felix Selgert\***  
Institut für Geschichtswissenschaft  
Universität Bonn

### Abstract

The correct detection of dense article layout and the recognition of characters in historical newspaper pages remains a challenging requirement for Natural Language Processing (NLP) and machine learning applications on historical newspapers in the field of digital history. Digital newspaper portals for historic Germany typically provide Optical Character Recognition (OCR) text, albeit of varying quality. Unfortunately, layout information is often missing, limiting this rich source's scope. Our dataset is designed to enable the training of layout and OCR models for historic German-language newspapers. The Chronicling Germany dataset contains 693 annotated historical newspaper pages from the time period between 1852 and 1924. The paper presents a processing pipeline and establishes baseline results on in- and out-of-domain test data using this pipeline. Both our dataset and the corresponding baseline code are freely available online. This work creates a starting point for future research in the field of digital history and historic German language newspaper processing. Furthermore, it provides the opportunity to study a low-resource task in computer vision.

# Article Separation

- Automatic newspaper article separation seems to be in reach

[Home](#) > [Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine](#)

[Collaboration](#) > Conference paper

## STRAS: A Semantic Textual-Cues Leveraged Rule-Based Approach for Article Separation in Historical Newspapers

Conference paper | First Online: 30 November 2023

pp 89–105 | [Cite this conference paper](#)



[Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration](#)

(ICADL 2023)

[Home](#) > [Linking Theory and Practice of Digital Libraries](#) > Conference paper

## LIAS: Layout Information-Based Article Separation in Historical Newspapers

Conference paper | First Online: 26 September 2024

pp 256–272 | [Cite this conference paper](#)



[Linking Theory and Practice of Digital Libraries](#)

(TPDL 2024)

# “Disneyland for Researchers”

- Dagstuhl Seminar 22292: Computational Approaches for Digitized Historical Newspapers (17–22 July 2022)
- Dagstuhl Infos and Report <https://www.dagstuhl.de/22292>

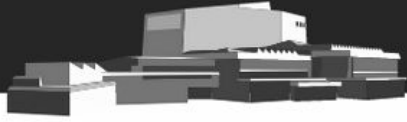


...in 10 years from today?



# Outlook

- Globally, hundreds of millions of newspaper pages are available digitally, with open licenses and programmatically retrievable by API
- Users can search content in text and image format
- Articles are structured as content units that are citable and referenceable
- Named Entities appearing in the content are identified, disambiguated and linked to authority sources
- Users can find semantically relevant content without requiring an exact keyword match using embeddings instead
- Content in languages unknown to the user can be translated on the fly
- Datasets bundle specific newspapers as corpora for research purposes
- A lightweight “Newspaper Network Notation Framework” (NNNF) guides standardised implementations of interoperable newspaper collections as data that are fit for computational use
- Is it time for another joint portal? Do we need it?



**Staatsbibliothek  
zu Berlin**  
Preußischer Kulturbesitz



# **Thank you for your attention! Questions?**

These slides online:

[\*\*http://sbb.berlin/54anj\*\*](http://sbb.berlin/54anj)

