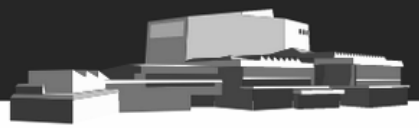


Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Beyond word clouds

Practical applications in challenging cultural contexts

Jana Götze | Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
CLASP Seminar | 10 June 2026
credit for many slides: Clemens Neudecker
jana.goetze@sbb.spk-berlin.de

Let’s start with a quiz

Fill the blank!

Berlin is the capital of _____

- FRG / Germany (1990 - now)
- GDR / East Germany (1949-1989)
- Soviet Occupation Zone (1945-1949)
- Third Reich (1933-1945)
- Weimar Republic (1918-1933)
- Kingdom of Prussia (1701-1918)
- Margraviate of Brandenburg (1618-1701)

Agenda

I. Background
II. NLP data, tasks and challenges
III. Related ongoing work
IV. Opportunities

Staatsbibliothek zu Berlin – Berlin State Library (SBB, “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
- 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 >240,000 digitized documents under Public Domain license
- **Stabi Lab** for experiments, events, datasets, digital humanities

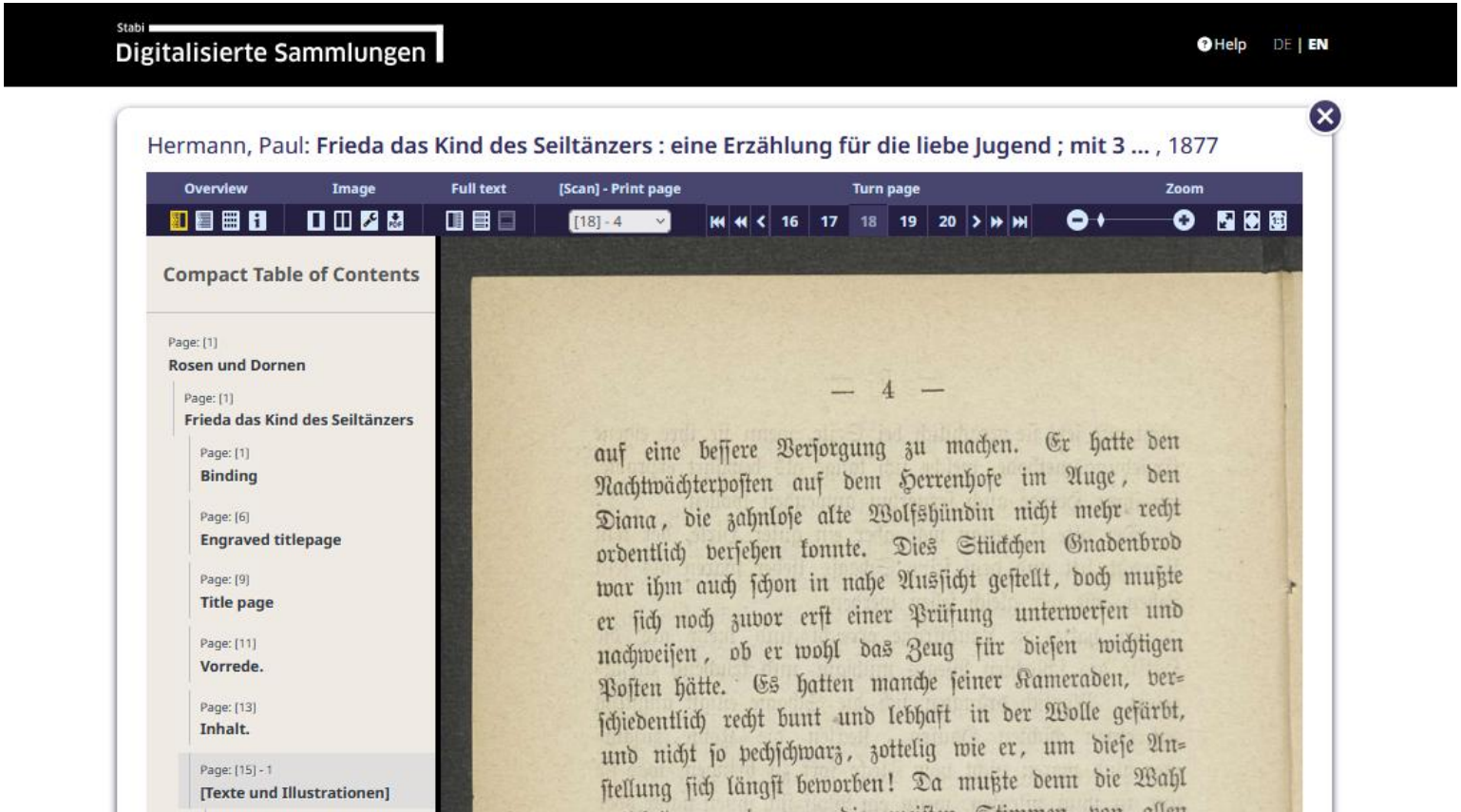
Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

What cultural heritage organisations do

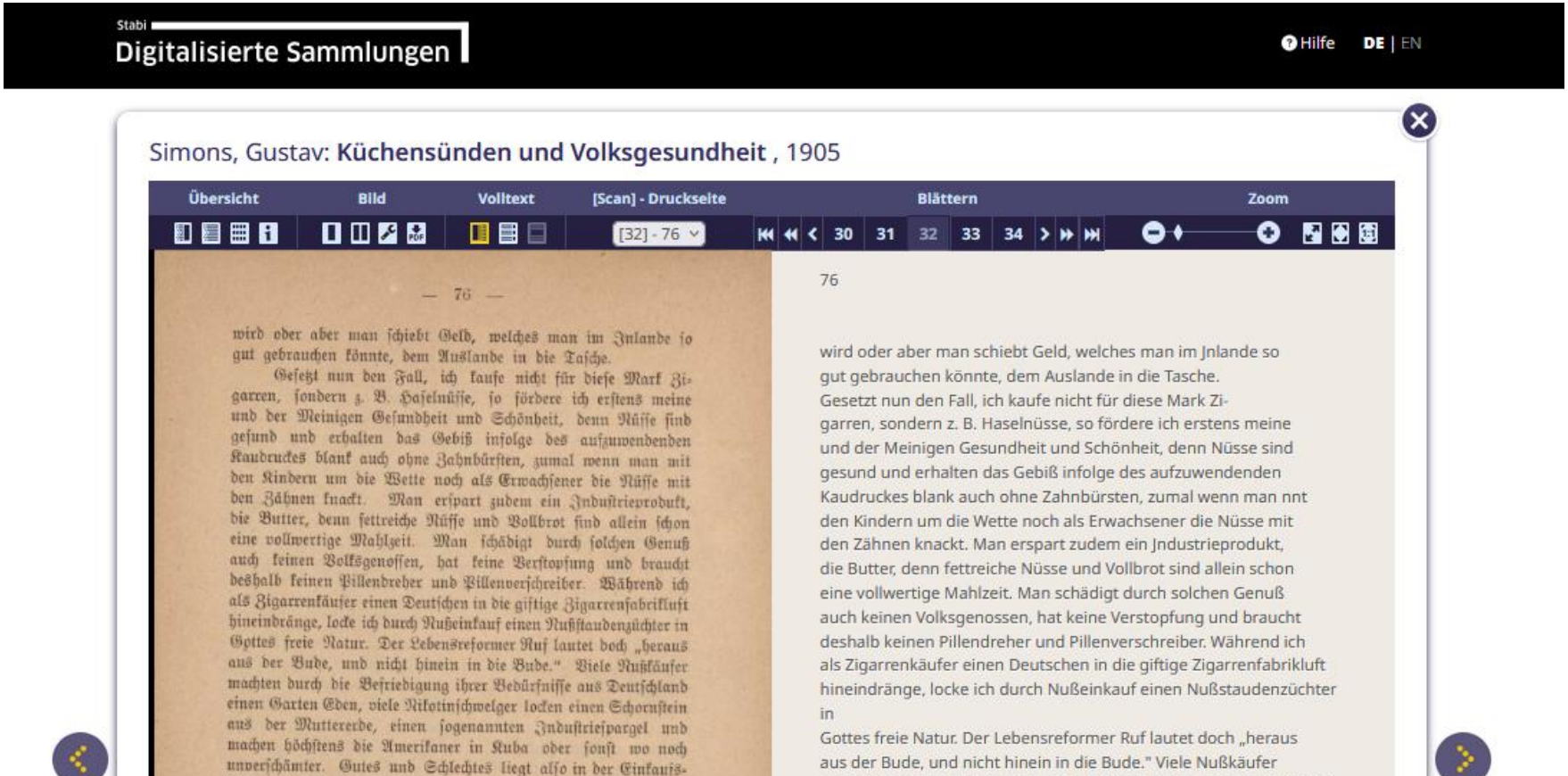
Provide data and services to the user

- Indexation, analysis and enhancement of digital data
- Create new services for users and scientists
- Growing (mostly open) data can serve as training data
- Expert knowledge about material and content creates quality data
- Sensitivity to data quality in creating, maintaining, and using data
- Transparency, data privacy and responsibility wrt AI are not just buzz words (public institution!)

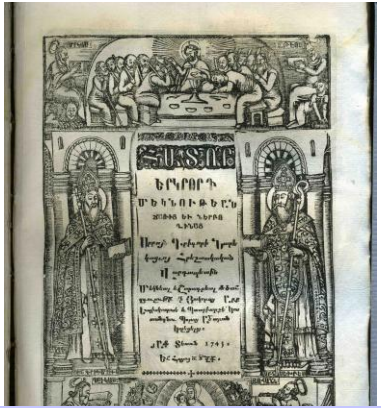
What digital access looks like



What digital access looks like

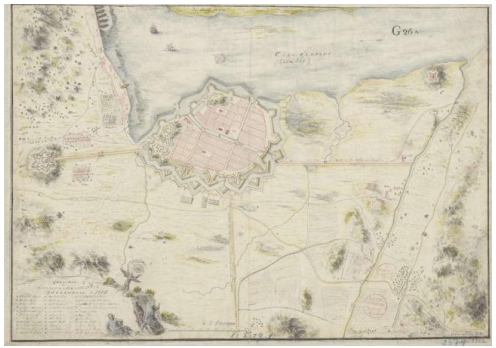


The variety of data is large



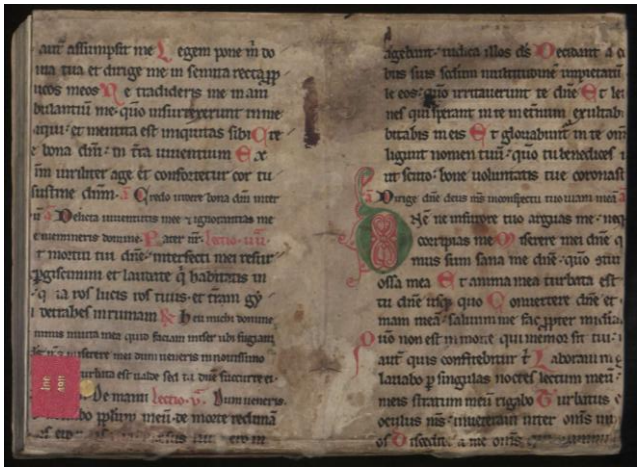
- occidental, oriental and East Asian manuscripts
- geographical maps, globes and atlases
- music scores
- autographs and bequests
- historical and modern printed matter
- 650,000+ newspaper issues
- 170,000+ photographs

- **historical data**
- **heterogeneous data**
- **a lot of data**

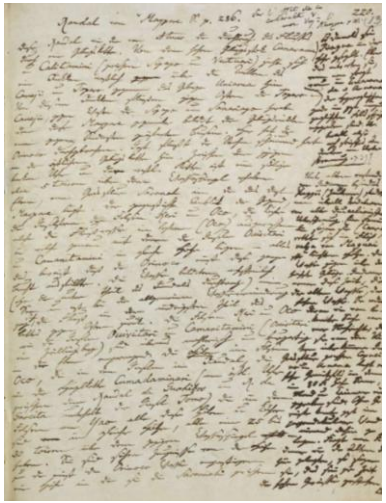


Digitized collections

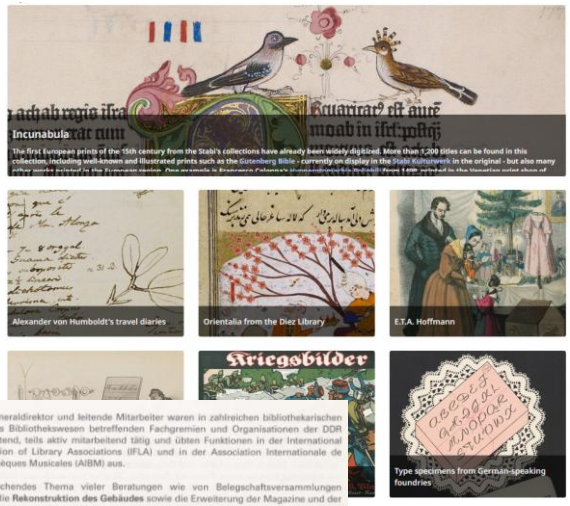
240,794 digitized objects



1477



1800



Der Generaldirektor und leitende Mitarbeiter waren in zahlreichen bibliothekarischen und das Bibliothekswesen betreffenden Fachgremien und Organisationen der DDR teils leitend, teils aktiv mitarbeitend tätig und übten Funktionen in der International Federation of Library Associations (IFLA) und in der Association Internationale de Bibliothèques Musicales (AIBM) aus.

Beherrschendes Thema vieler Beratungen wie von Belegschaftsversammlungen waren die **Rekonstruktion des Gebäudes** sowie die Erweiterung der Magazine und der Lesesaalplätze. Im Mai 1976 begannen Instandsetzungsarbeiten an den Fassaden und zwar zunächst an der Front der Universitätsstraße. Sie wurden in der Folgezeit unkontinuierlich weitergeführt.

Die überaus angespannte Platzsituation in den Hauptmagazinen führte zu einer Reihe von Notmaßnahmen und zur Suche nach Räumlichkeiten für die Auslagerung von Buchbeständen.

Nachdem im April 1975 die stählerne Dachkonstruktion über der Ruine des alten Kuppellesesaals aus Sicherheitsgründen gesprengt werden mußte und Abbruch- und Transportarbeiten einsetzten, die für viele der Mitarbeiter Beeinträchtigungen durch Lärm und Staub brachten, konnten schließlich grundlegende Entscheidungen getroffen werden. Die neue Variante, Ergebnis jahrelanger Diskussionen, Entwürfe und Kapazitätsberechnungen, sah die Errichtung von vier Büchertürmen mit einer Nutzfläche für etwa 2,4 Millionen Bände und eines Zwischenbaues vor sowie ein zusätzliches Funktionsgebäude außerhalb der Bibliothek. Der Baubeginn sollte sich aber noch weitere Jahre hinziehen.[1]

1.2. Wissenschaftliches Leben

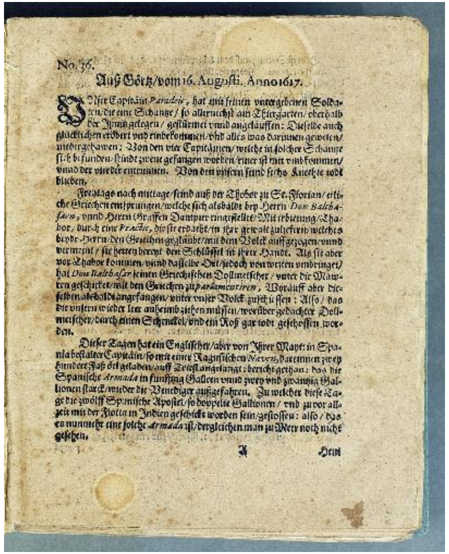
Mitarbeiter der DSB haben durch Forschungsarbeiten, fachwissenschaftliche und populärwissenschaftliche Veröffentlichungen sowie Vorträge bei Veranstaltungen oder auf Kongressen im In- und Ausland wesentlich zur Förderung des geistigen und kulturellen Lebens beigetragen und Arbeitsergebnisse und Bestände der Bibliothek für Fachkollegen und die Öffentlichkeit erschlossen.

Stellvertretend für viele dieser Beiträge seien die Mitarbeit an der Festschrift „Studien zur Buch- und Bibliotheksgeschichte. Hans Lülfing zum 70.Geburtstag" erwähnt, das Handschrifteninventar „Der Nachlaß Hans Delbrück", die Publikation „America in maps" oder die Veröffentlichung von Band 7 der Konversationshefte Ludwig van Beethovens.

Weiten Widerhall fand auch die **Konferenz aus Anlaß des 75jährigen Bestehens der Arbeit am Gesamtkatalog der Wiegendrucke** im November 1979. Fünf von den Referaten, die Forschungsergebnisse zur Buchdruck-, Wissenschafts- und Kulturgeschichte

2001

Newspapers (ZEFYS portal)


No. 36.
Auß Görtz/vom 16. Augusti. Anno 1617.

1617


Berlinische Nachrichten
Von Staats- und Gelehrten Sachen.
No. 80. Donnerstags den 4. Julius 1776.

Berlinische Nachrichten von Staats- und gelehrten Sachen
1776


المؤيد

al- Mu'aiyad
1897

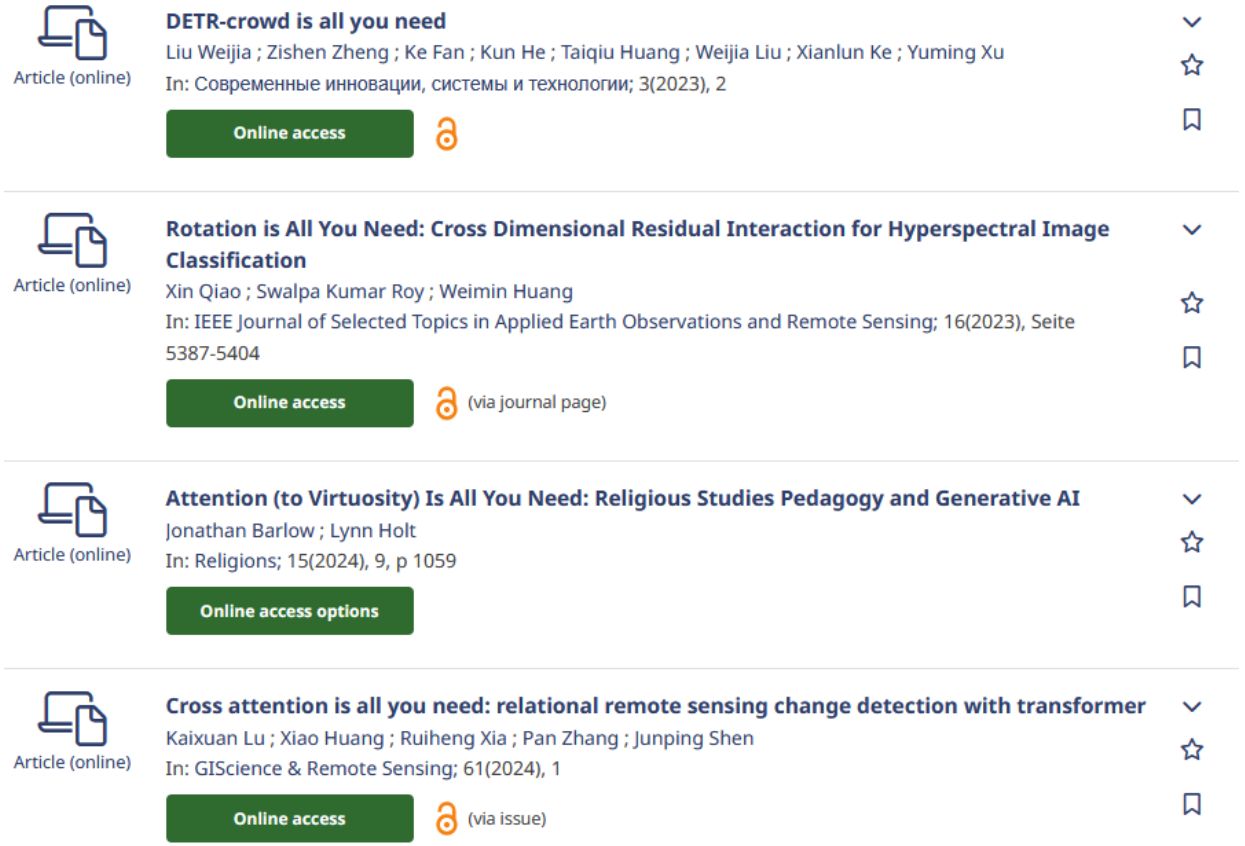

DZIENNIK BERLINSKI
wychodzi codziennie
Rok XLV Piątek, dnia 31 marca 1939 r. Nr. 76
Wspólnym wysiłkiem zwiększa Polska potęgę swego lotnictwa
„Jednomyślność poglądów narodu polskiego"
Koniec wojny hiszpańskiej

Dziennik Berliński
1939

Stabi has 4 goals to serve the user (with ML methods)

- Provide **search and retrieval** across all collections, for text and image content
- Research and develop **open source technologies** for cultural heritage
- Provide collections as **open and machine-readable** digital data (“Collections as Data”)
- **Responsibly use AI** and curate digital data under consideration of problematic content based on ethical, legal and social criteria

There is also contemporary material



Article (online)

DETR-crowd is all you need

Liu Weijia ; Zishen Zheng ; Ke Fan ; Kun He ; Taiqiu Huang ; Weijia Liu ; Xianlun Ke ; Yuming Xu

In: Современные инновации, системы и технологии; 3(2023), 2

Online access

Article (online)

Rotation is All You Need: Cross Dimensional Residual Interaction for Hyperspectral Image Classification

Xin Qiao ; Swalpa Kumar Roy ; Weimin Huang

In: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing; 16(2023), Seite 5387-5404

Online access (via journal page)

Article (online)

Attention (to Virtuosity) Is All You Need: Religious Studies Pedagogy and Generative AI

Jonathan Barlow ; Lynn Holt

In: Religions; 15(2024), 9, p 1059

Online access options

Article (online)

Cross attention is all you need: relational remote sensing change detection with transformer

Kaixuan Lu ; Xiao Huang ; Ruiheng Xia ; Pan Zhang ; Junping Shen

In: GIScience & Remote Sensing; 61(2024), 1

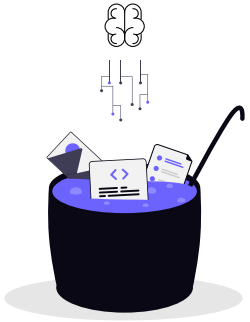
Online access (via issue)

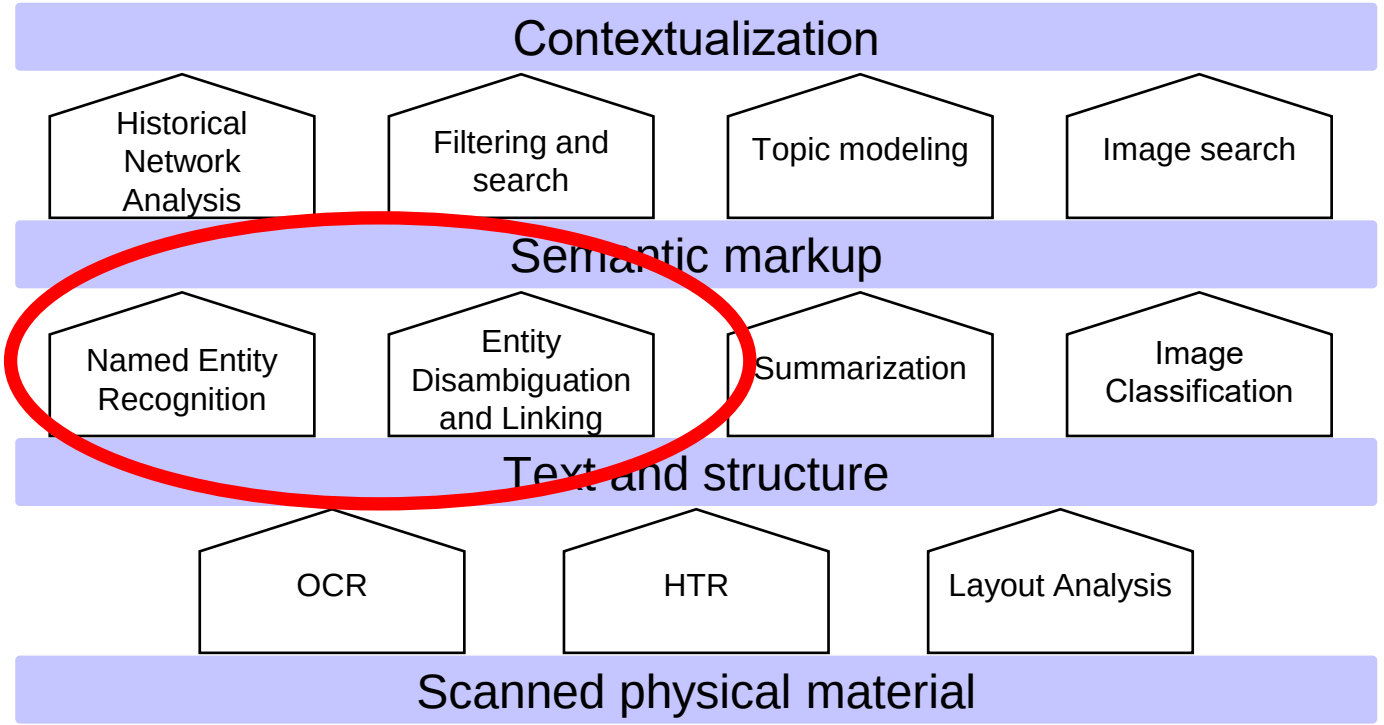
The main target is the digital humanities researcher

- Modern methods vs. historic material
  - models and methods must work on a variety of material (type, age, quality)
- Results as a means to an end
  - search and filtering
  - topic modeling
  - network analysis
- Deployment
  - usable tools, incl. frontend
  - must process a huge volume of data
  - public institution has limited infrastructure (finances, maintaining, hardware)

Discovery vs. modeling

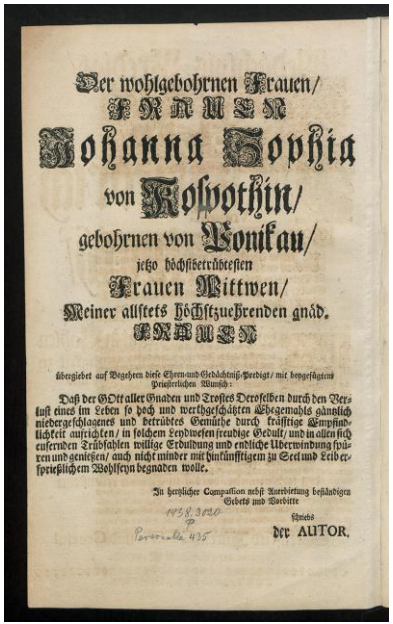
From scanned image to contextualized information

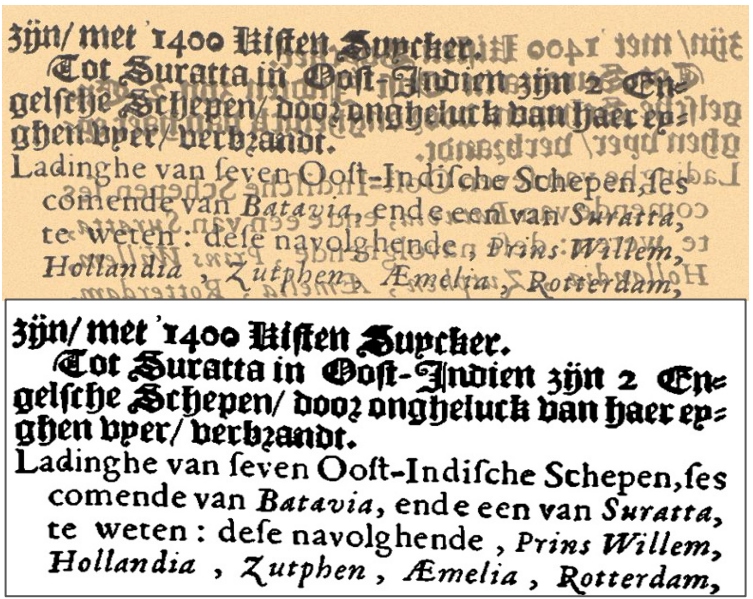




Prep 1: Image Enhancement and Binarization

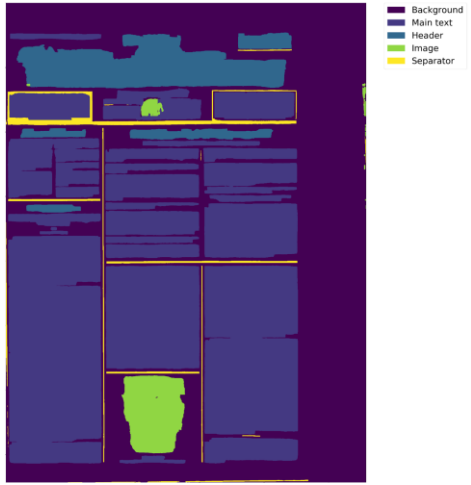
A hybrid CNN-Transformer model for Historical Document Image Binarization, 2023. https://doi.org/10.1145/3604951.3605508 | https://github.com/qurator-spk/sbb_binarization

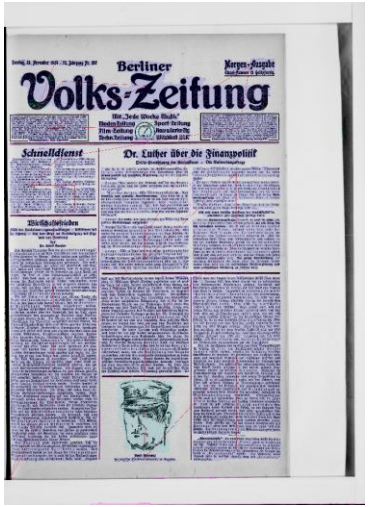




Prep 2: Document Layout Analysis (Segmentation)

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

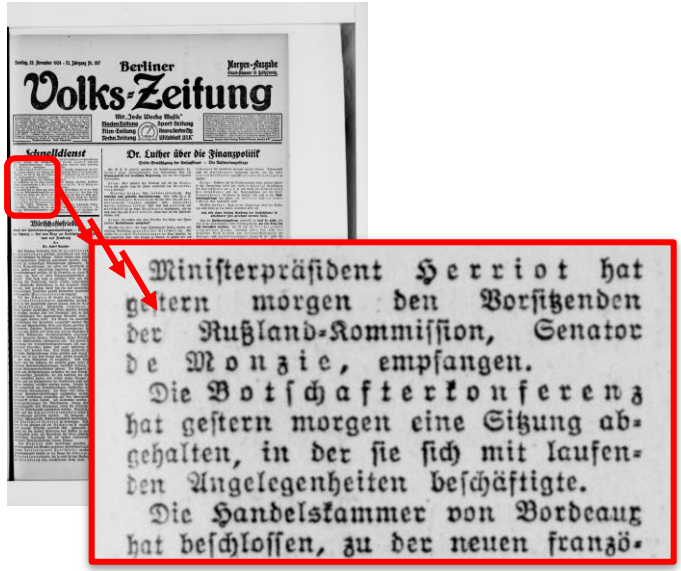




Prep 3: Text Recognition (OCR, HTR)

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 Hcrri 0 t hat Die Feierlichkeiten zur UcbcrfUhrung geilern morgen de» Vorsitzenden der Autzlanü-Komniission, Cenator dl M 0 II z i e, empfangen. Die Botschafterkonferenz bat gestern morgen eine Citzung ab, gehalten, in der sie sich mit laufenten Angelegenheiten beschäftigte. Die Handelskammer von Bordeaux bat defchloffen, zu der neuen franzö_ sischen Inlandsanleihe 1 Million Francs zu zeichne». Aus Genf sind in Sofia zwei Delegierte der Bälkerdundskommifsion zur Prüfung der Frage der Massen, auswanderung der bul. garifchen Bevölkerung aus Thrazien und Mazedonien und der Ictzien Beschwerde der bulgarischen Regierung an die zu ständige Dölkerbundskommisston eingetroffen.

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 laufenden Angelegenheiten beſchäftigte. Die Handelskammer von Bordeaux hat beſchloſſen, zu der neuen franzöſiſchen Jnlandsanleihe 1 Million Francs zu zeichnen. Aus Genf ſind in Sofia zwei Delegierte der Völkerbundskommiſſion zur Prüfung der Frage der Maſſenauswanderung der bulgariſchen Bevölkerung aus Thrazien und Mazedonien und der letzten Beſchwerde der bulgariſchen Regierung an die zu– ſtändige Völkerbundskommiſſion eingetroffen.

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://github.com/qurator-spk/sbb_ner | https://github.com/qurator-spk/sbb_ned

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

?

Édouard_Herriot
Q274344 (0.75)

Russland
Q159 (0.27)

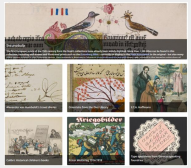
Bordeaux
Q1479 (0.58)

Named Entity Recognition and Linking Data

- mostly newspaper data, German or multilingual (de/fr/en)
- historical vs. contemporary

Unlabeled data

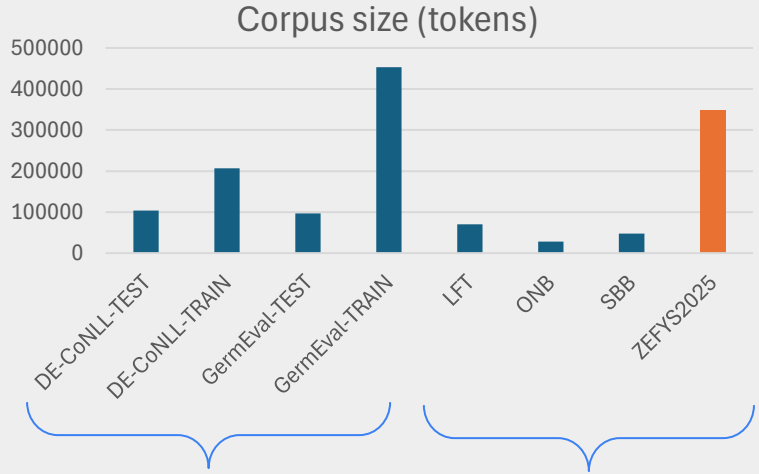
- DC-SBB: Digitized Collections
- 1470 – 1945
- 2,333,647 pages



Knowledge bases

- Wikidata
- Wikipedia
- GND – Integrated Authority File

Annotated data



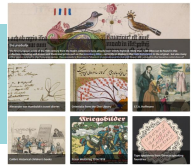
- CoNLL 2003, Frankfurter Rundschau 1992
- German Wikipedia and various online newspapers

- LFT, 1926
- ONB, 1710 – 1873
- SBB, 1872 – 1930
- ZEFYS, 1830 – 1940

1) Named Entity Recognition

Multilingual BERT-Base model

+unsupervised pre-training on historic text



- How well do models perform on historic text when they are trained on contemporary data?
  - Do we need historic ground truth?

A lot worse than on the contemporary data.

- Does unsupervised training on historic text help?

Yes, for the historical datasets, results improve irrespective of the ground truth.

BERT models outperform previous biLSTM+CRF results but cannot reach best results

1) Named Entity Recognition numbers

2019

	pre-train:	none	DC-SBB
train	test	F_1	F_1
GermEval + CoNLL	CoNLL	**80.2**	79.4
	GermEval	**88.0**	85.7
	LFT	55.1	**55.2**
	ONB	58.6	**60.1**
	SBB	64.1	**65.1**

gual-cased (Riedl and Padó, 2018)

5-fold cross validation on	pre-train	precision	recall	F_1	F_1
ONB	Newspaper (1703-1875)	-	-	-	-
	DC-SBB+GermEval + CoNLL	81.5 ±1.8	87.8 ±1.4	84.6 ±1.5	-
	DC-SBB + GermEval	81.6 ±2.5	87.5 ±1.6	84.5 ±1.8	-
	DC-SBB + CoNLL	81.7 ±2.8	87.5 ±1.9	84.5 ±2.3	-
	DC-SBB	81.8 ±2.3	87.1 ±2.1	84.3 ±2.0	-
	GermEval	80.8 ±2.1	85.4 ±1.2	83.0 ±1.4	78.56
	GermEval + CoNLL	80.0 ±1.5	84.7 ±1.6	82.3 ±1.5	-
	CoNLL	79.1 ±2.5	84.5 ±2.1	81.7 ±2.2	76.17
	none	78.0 ±2.4	84.1 ±1.9	80.9 ±2.0	73.31

2025

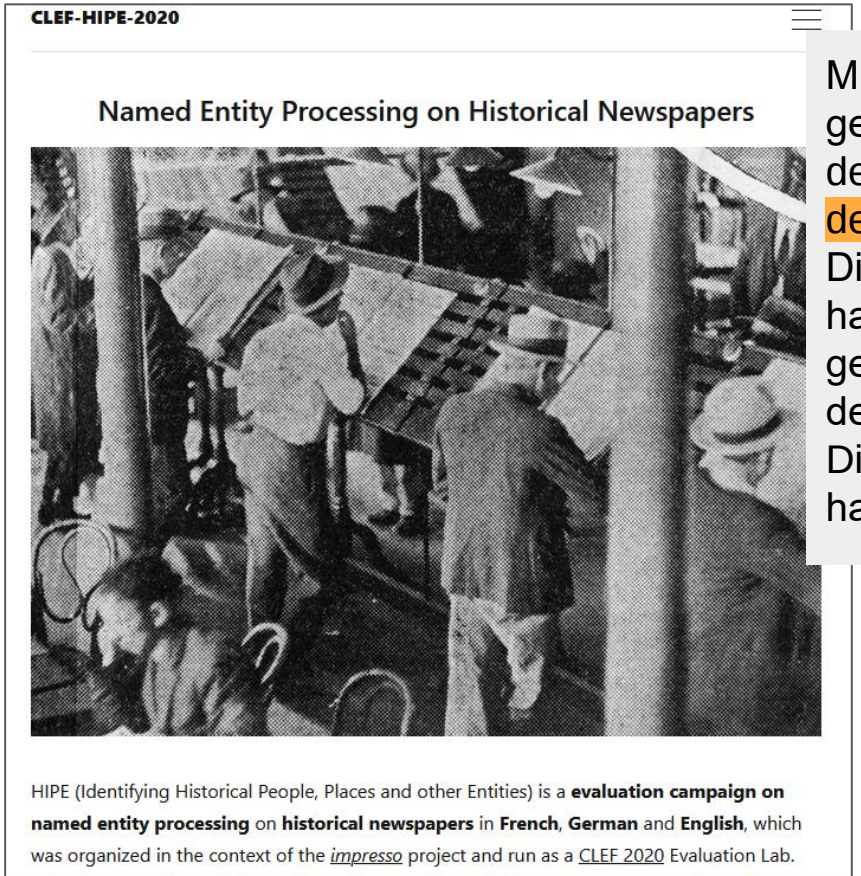
model	F_1^{TE}	$F_{1_{PZ}}^{TE}$	$F_{1_{PH}}^{TE}$	F_1^{PER}	F_1^{LOC}	F_1^{ORG}
			Europeana-ÖNB			
Europ-ELECTRA	0.84	0.86	**0.88**	0.82	0.94	0.44
XLM-RoBERTa	0.84	**0.88**	0.86	0.85	0.92	0.13
hmBERT	0.84	**0.86**	**0.86**	0.78	0.92	0.50
RoBERTa	0.81	0.84	**0.87**	0.81	0.92	0.20
GBERT	0.83	**0.84**	0.83	0.75	0.92	0.40
GermanBERT	0.82	0.83	**0.84**	0.73	0.92	0.53
hmBERT-mini	0.70	0.76	**0.78**	0.68	0.86	0.28
DistilBERT	0.70	0.74	**0.75**	0.61	0.86	0.15
hmBERT-tiny	0.50	0.63	**0.65**	0.48	0.78	0.38

Training on the new larger ZEFYS dataset

Labusch, K. et al. (2019). *BERT for Named Entity Recognition in Contemporary and Historical German*. KONVENS 2019.

Schneider, S. et al. (2025). *ZEFYS2025: A German Historical Newspaper Dataset for Named Entity Recognition and Entity Linking*. KONVENS 2025.

2) Linking Entities



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

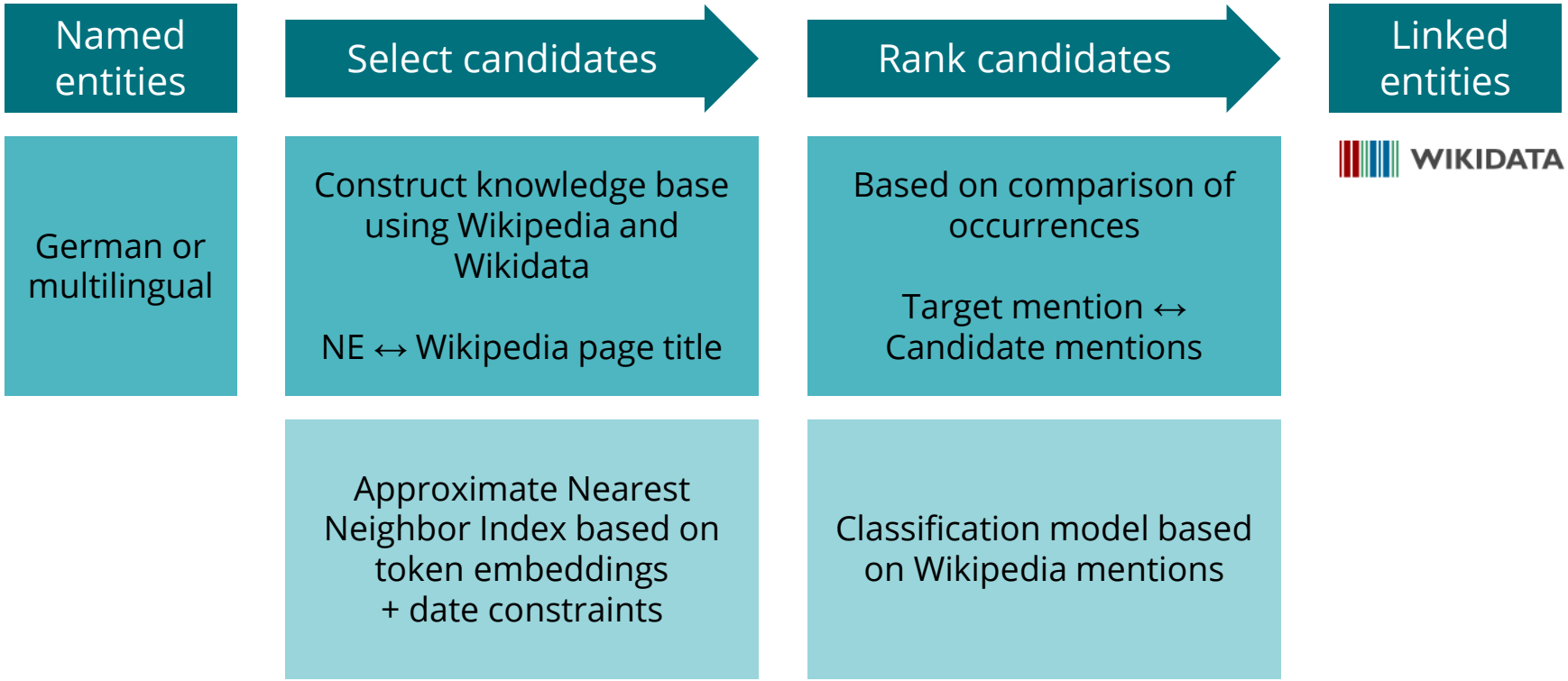
?

Édouard_Herriot
Q274344 (0.75)

Russland
Q159 (0.27)

Bordeaux
Q1479
(0.58)

2) Linking Entities, continue with BERT models

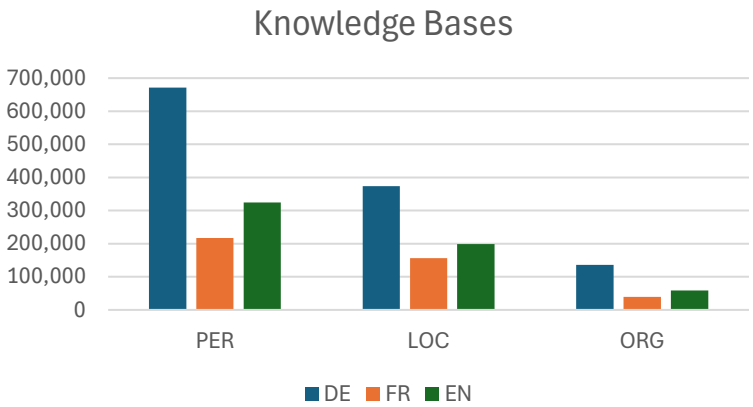


Labusch, K., & Neudecker, C. (2020). *Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT*. CLEF 2020.

Labusch, K., & Neudecker, C. (2022). *Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT*. CLEF 2022.

2) Named Entity Linking results

- How well do models perform on historic text and for the different languages?



Results are similar for FR and DE even though the knowledge bases have different sizes (coverage of the test data is similar)

Lower scores for EN → OCR

	2020		2022	
	SBB	Best system	SBB	Best system
Language	F1	F1	F1	F1
German	0.389	0.534	0.506	0.506
French	0.407	0.598	0.596	0.602
English	0.141	0.531	0.393	0.546

Effect of knowledge base size and quality is not clear!

coverage
.86 – .99

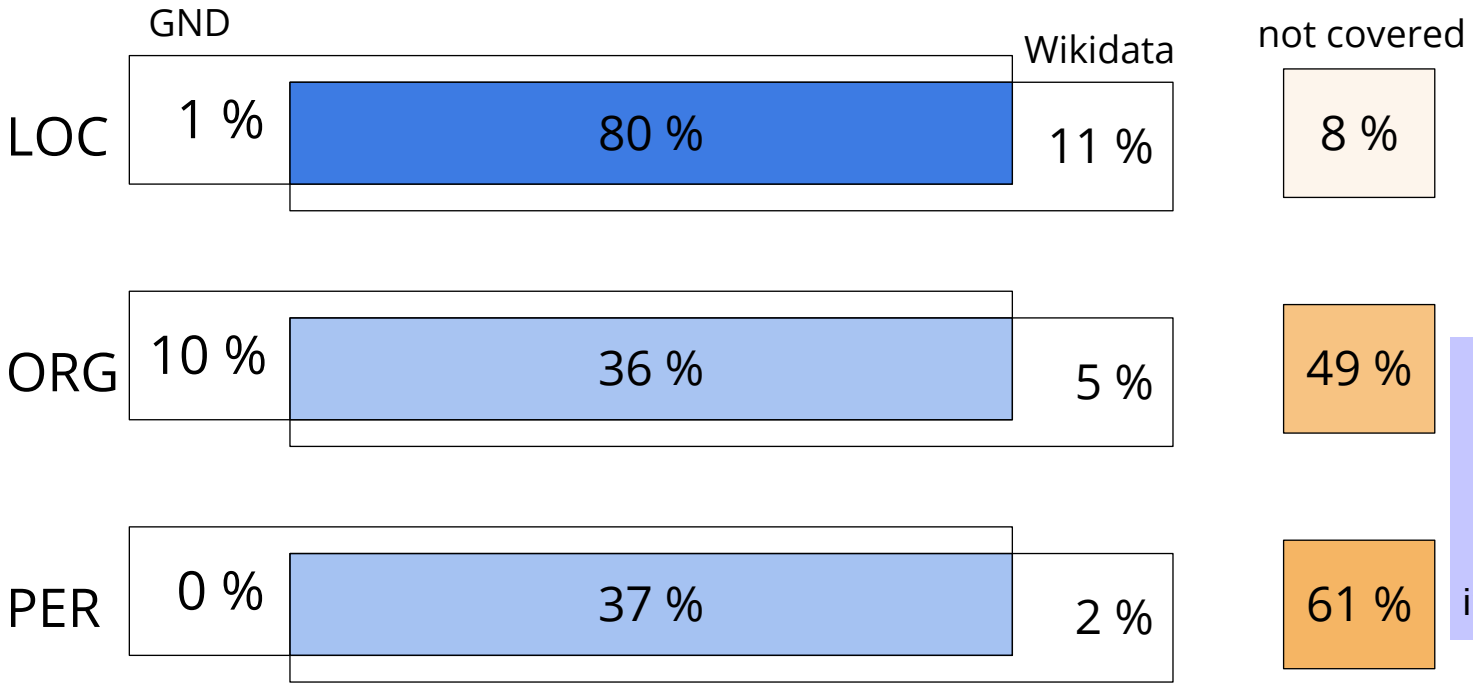
Labusch, K., & Neudecker, C. (2020). *Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT*. CLEF 2020.

Labusch, K., & Neudecker, C. (2022). *Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT*. CLEF 2022.

Contemporary knowledge bases cover the same entities

Overall, GND covers 68.1% / 51.8%, Wikidata covers 69.5% / 54.1%

N = 791 entities



many ORG and PER that are mentioned in historical newspapers are insignificant today

Knowledge base quality matters, too

- **duplicate entries**
- **missing entities (mismatching focus)**
  - GND focusses on authors, newspapers on public persons
- **incomplete entries**
  - “empty“ entries contain attributes that are matched by the method
- missing **spelling variants**
  - newspaper used a spelling that is missing from the correct entry
- entities with **changing extension**

Reichstag (Q160208)
parliament of Germany from 1871 to 1918
16 statements, 25 sitelinks – 02:53, 25 May 2026

Reichstag (Q321246)
parliament of the North German Confederation 1867-1870
14 statements, 11 sitelinks – 16:29, 27 January 2026

Reichstag (Q878525)
parliament of the Third Reich from 1933 to 1945
13 statements, 21 sitelinks – 12:14, 18 May 2026

Do we choose the most extensive one? the one closest in time? only the one that matches in every detail?

NER and EL takeaways

- **Longterm goal and requirement:**
  - deliver robust and solid baseline performance even without optimization on the diverse material (age, heterogeneity, size)

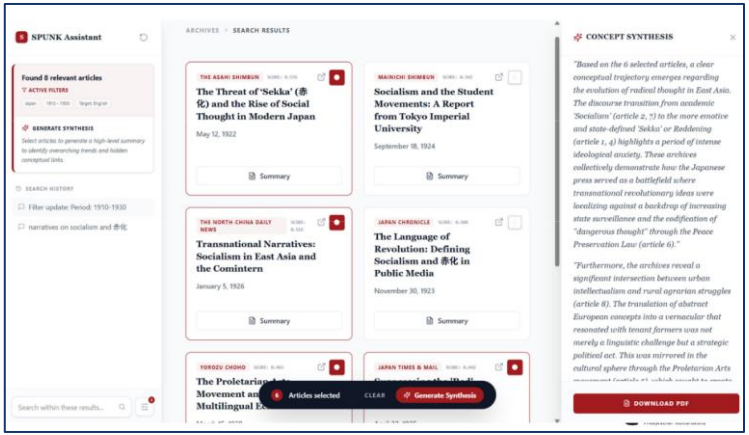
- OCR is crucial
- NER results are stable, EL results are a starting point
- improve knowledge base construction
- exploit more constraints like time/date, e.g. publication dates

Labusch, K., & Neudecker, C. (2022). *Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT*. CLEF 2022.

There are more NLP tasks

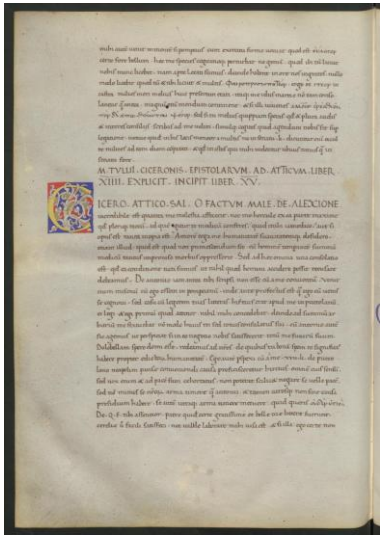
Multilingual search and summarization

Multilingual search → Summarization → Translation



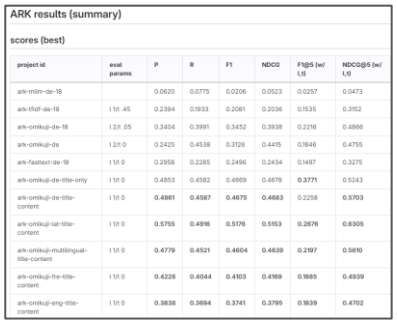
Hand-written text processing

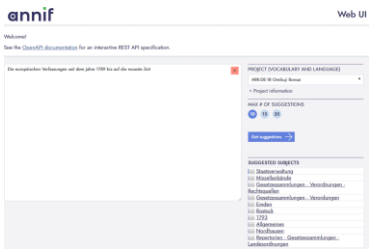
- Text initials
- Finding duplicates
- Identify texts



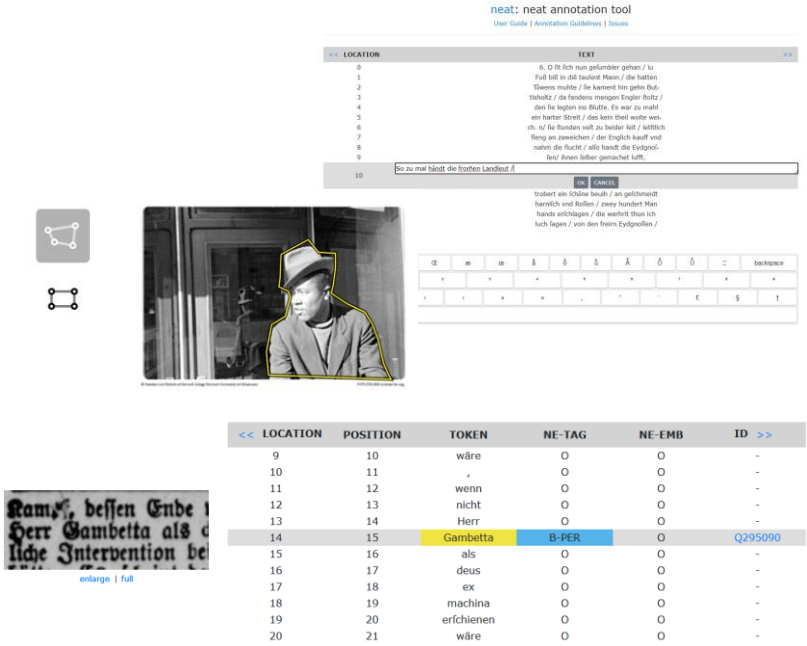
Semi-automatic Subject Indexing

- Analyse requirements in different subject domains at Stabi
- Train and evaluate Annif models for ARK and BK





Transcription and annotation



https://github.com/qurator-spk/neat

Image (Similarity) Search

- Develop a multi-modal text-image-search
  - Find images by upload of an example image
  - Find images by a textual description (48 languages currently supported)
  - Find all images within a specific document

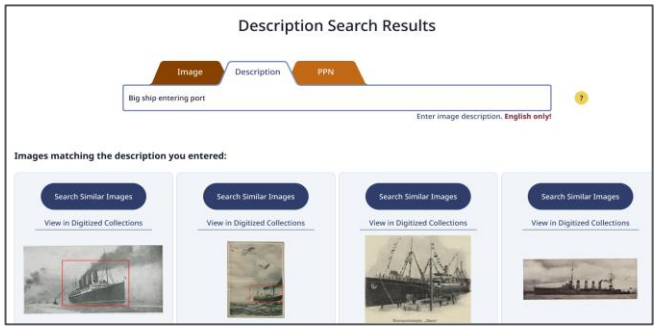
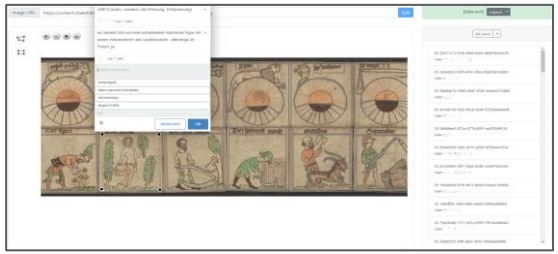


Image Analysis for Special Subject Domains

- Develop web-based image annotation tool
- Adapt and train AI models specifically for
  - Iconographic images (ICONCLASS)
  - Provenance features (e.g. printers marks)
  - Watermarks
  - Illustrations in children's books



And even more challenges

- **Ethically, socially or legally problematic content** (such as colonialism, National Socialism or the underrepresentation of marginalised groups) in the source data must be identified and treated and contextualised with appropriate caution and care
    - Datasheets for Digital Cultural Heritage
- **Newer material is subject to a number of laws** that prevent their online publication.
    - Formally describe access constraints as metadata
    - Develop infrastructure to enable access
    - Work with derivative formats
- **Public infrastructure is sensitive to bot attacks and data scraping**

But there are also great opportunities in using ML methods

develop methods for more efficient processing

create more ways of accessing data for the user

model data that changes with time and context

use diverse and trustworthy data

With thanks to Steven Claeyssens for this excellent meme!

Where to go from here

- multilinguality: SBB has documents in many languages
  - *search/retrieval, summarization*
- entity linking: enable historic research through entity search
  - *search for entities by ID*
- image processing
  - *search, comparison*
- derived text formats
  - *for material that cannot be accessed directly*
- pre-training with more rich, varied and contextualized data

Outputs

- Open Source Software **github.com/qurator-spk**
  - Document Layout Analysis (Segmentation)
  - Optical Character Recognition
  - Image Similarity Search
  - Named Entity Recognition
  - Entity Disambiguation and Linking
- Models **huggingface.co/SBB**
  - Document Layout Analysis
  - Named Entity Recognition
  - Entity Linking
- Datasets
  - Hugging Face **huggingface.co/SBB**
  - Zenodo **zenodo.org/communities/stabi/**

OCR data

Datasets of Staatsbibliothek zu Berlin - Berlin State Library

Published June 26, 2019 | Version 1.0

Dataset Open

OCR fulltexts of the Digital Collections of the Berlin State Library (DC-SBB)

Labusch, Kai[1] ; Zellhöfer, David[1]

Show affiliations

The digital collections of the SBB contain 153,942 digitized works from the time period of 1470 to 1945.

At the time of publication, 28,909 works have been OCR-processed resulting in 4,988,099 full-text pages.
For each page with OCR text, the language has been determined by *langid* (Lui/Baldwin 2012).

corpus-entropy.pkl	entropy rate per document page
corpus-language.pkl	language per document page
corpus.zip	fulltext corpus (extracts to .txt format)
de_corpus.zip	German sub-corpus (extracts to .txt format)
selection_de.pkl	Selection list of German documents
xml2csv_alto.csv	fulltext corpus per document page (incl.OCR word confidences)

Sources

Marco Lui and Timothy Baldwin. 2012. Langid.py:

An off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations,

ACL '12, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics

https://zenodo.org/records/3257041

NER models

README Apache-2.0 license

NER - Demo

Task: Named Entity Recognition

Model: DC-SBB + CONLL + GERMEVAL

Input text: Die Staatsbibliothek zu Berlin (ab 1661: Churfürstliche Bibliothek; ab 1701: Königliche Bibliothek; ab 1918: Preußische Staatsbibliothek; ab 1954: Deutsche Staatsbibliothek)[2] ist eine Einrichtung der Stiftung Preußischer Kulturbesitz, einer durch Bundesgesetz errichteten rechtsfähigen Stiftung des öffentlichen Rechts mit Sitz in Berlin. Die Bibliothek sammelt für den Spitzenbedarf der Forschung wissenschaftlich relevante Literatur aus allen Zeiten, allen Ländern und in allen Sprachen. Sie ist eine der größten Bibliotheken Deutschlands und darüber hinaus eine der größten der Erde

Go

Ergebnis:

Die Staatsbibliothek zu Berlin (ab 1661 : Churfürstliche Bibliothek ; ab 1701 : Königliche Bibliothek ; ab 1918 : Preußische Staatsbibliothek ; ab 1954 : Deutsche Staatsbibliothek) [2] ist eine Einrichtung der Stiftung Preußischer Kulturbesitz , einer durch Bundesgesetz errichteten rechtsfähigen Stiftung des öffentlichen Rechts mit Sitz in Berlin .
Die Bibliothek sammelt für den Spitzenbedarf der Forschung wissenschaftlich relevante Literatur aus allen Zeiten , allen Ländern und in allen Sprachen .
Sie ist eine der größten Bibliotheken Deutschlands und darüber hinaus eine der größten der Erde .
Die Staatsbibliothek zu Berlin ist die größte wissenschaftliche Universalbibliothek im deutschen Sprachraum .
Zu den bedeutendsten Unterstützern der Bibliothek gehört die Deutsche Forschungsgemeinschaft (DFG) mit Sitz in Bonn .

Legende:
[Person]
[Ort]
[Organisation]
[keine Named Entity]

Contributors 3

labusch Kai Labusch

cneud Clemens Neudecker

joergleh joergleh

Languages

Python 73.3% Makefile 19.9%
JavaScript 4.4% HTML 2.0%
Dockerfile 0.4%

https://github.com/qurator-spk/sbb_ner

API access

digital.staatsbibliothek-berlin.de

OAI 2.0 Request Results

More information to this OAI interface and the sets are available in the Lab of the SBB: https://lab.sbb.berlin/dc

Identify | ListRecords (oai_dc) | ListRecords (mets) | ListSets | ListMetadataFormats | ListIdentifiers

You are viewing an HTML version of the XML OAI response. To see the underlying XML use your web browsers view source option. More information about this XSLT is at the bottom of the page.

Datestamp of response	2026-05-23T15:20:13Z
Request URL	https://oai.sbb.berlin/

OAI Error(s)

The request could not be completed due to the following error or errors.

Error Code	badVerb

Required verb argument not found.

Identify | ListRecords (oai_dc) | ListRecords (mets) | ListSets | ListMetadataFormats | ListIdentifiers

About the XSLT

An XSLT file has converted the OAI-PMH 2.0 responses into XHTML which looks nice in a browser which supports XSLT such as Mozilla, Firebird and Internet Explorer. The XSLT file was created by Christopher Gutteridge at the University of Southampton as part of the GNU EPrints system, and is freely redistributable under the GPL.

If you want to use the XSL file on your own OAI interface you may but due to the way XSLT works you must install the XSL file on the same server as the OAI script, you can't just link to this copy.

For more information or to download the XSL file please see the OAI to XHTML XSLT homepage.

https://oai.sbb.berlin/oai

Summary

Libraries have real data and real users

Curated data that is not available elsewhere
+ expert knowledge

The material is a challenge at every level of processing (age, heterogeneity, size)

Libraries can be a testbed for NLP methods

Ideas and suggestions welcome!

References

Illustrations: Katerina Limpitsouni / undraw.co

Alkemade, H. et al. (2023). *Datasheets for Digital Cultural Heritage Datasets*.

Labusch, K., & Neudecker, C. (2020). *Named Entity Disambiguation and Linking on Historic Newspaper OCR with BERT*. CLEF 2020.

Labusch, K., & Neudecker, C. (2022). *Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT*. CLEF 2022.

Labusch, K., & Neudecker, C. (2023). *Gauging the Limitations of Natural Language Supervised Text-Image Metrics Learning by Iconclass Visual Concepts*. HIP '23.

Labusch, K. et al. (2024). *Automatisierte semantische Anreicherung von historischen Texten*. b.i.t.online, 27(3), 232–241.

Labusch, K. et al. (2019). *BERT for Named Entity Recognition in Contemporary and Historical German*. KONVENS 2019.

Neudecker, C. et al. (2021). *A survey of OCR evaluation tools and metrics*. The 6th International Workshop on Historical Document Imaging and Processing, 13–18.

Neudecker, C. et al. (2021). *Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten*. Qualität in der Inhaltserschließung.

Neudecker, C. (2022*). Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries*. Proc. of the Third Conference on Digital Curation Technologies.

Menzel, S. et al. (2021). *Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten*. Qualität in der Inhaltserschließung.

Rezanezhad, V. et al. (2023). *A hybrid CNN-Transformer model for Historical Document Image Binarization*. HIP '23.

Rezanezhad, V. et al. (2023*). Document Layout Analysis with Deep Learning and Heuristics*. HIP '23.

Schaefer, R., & Neudecker, C. (2020). *A Two-Step Approach for Automatic OCR Post-Correction*. In Proc. of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.

Schneider, S. et al. (2025). *ZEFYS2025: A German Historical Newspaper Dataset for Named Entity Recognition and Entity Linking*. KONVENS 2025.

Zellhöfer, D. (2019). *Multimodal Datasets of the Berlin State Library*. In Proc. of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK).