

# Digitalisate kuratieren mit KI - von unstrukturierten Daten zu strukturierten Inhalten

Clemens Neudecker ([@cneudecker](#))

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Fachtagung Netzwerk maschinelle Verfahren in der Erschließung

10.-11. Oktober 2019, DNB, Frankfurt am Main

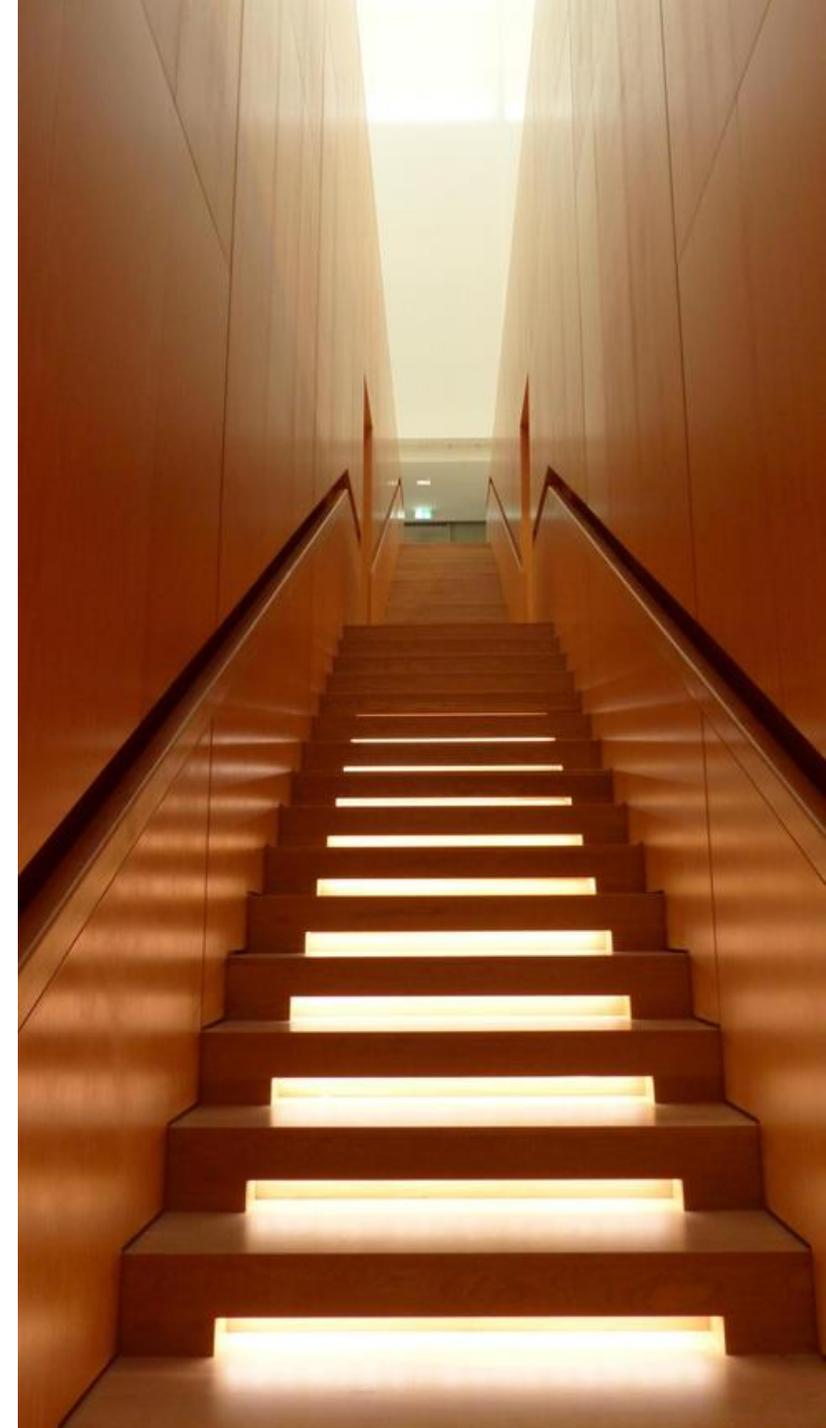


Staatsbibliothek  
zu Berlin  
Preußischer Kulturbesitz

**Qurator**  
Curation Technologies

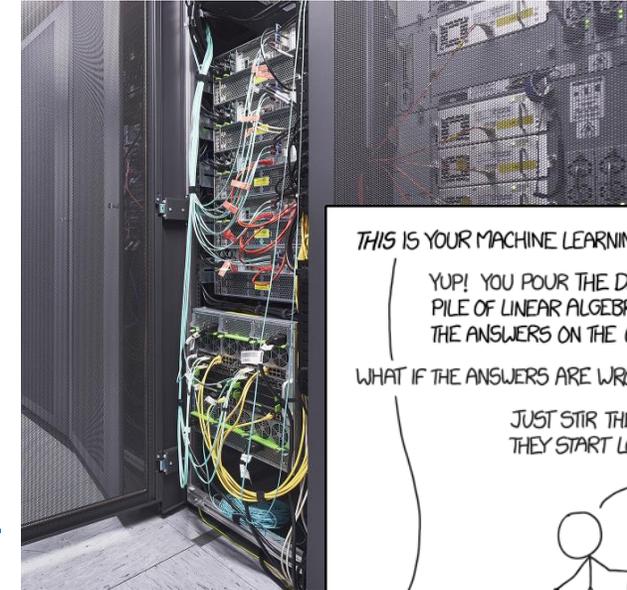
# Hintergrund

- > 12 Mio. Dokumente
- Metadaten (METS, MODS)
- Digitalisierte Sammlungen
  - <https://digital.staatsbibliothek-berlin.de/>
  - ca. 160,000 Digitalisate
  - ca. 5 Mio. Seiten OCR
- Digitalisierte Zeitungen
  - <http://zefys.staatsbibliothek-berlin.de/>
  - ca. 7 Mio. Seiten digitalisiert
  - ca. 3 Mio. Seiten OCR
- ca. 2,5 PetaBytes Daten



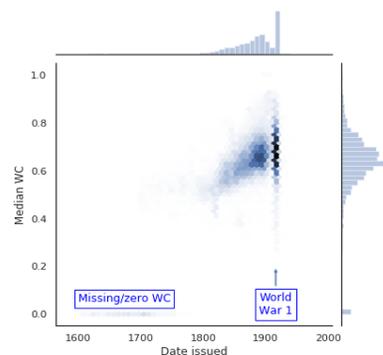
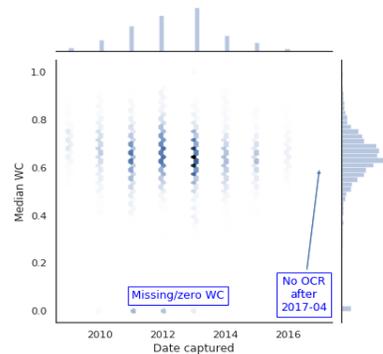
# QURATOR @ SBB

- Gemeinsame Projektsteuerung durch Forschungsreferent und Referatsleiter IT-Anwendungen und Datenmanagement
- 3x FTE Entwicklerstellen E13 FuE (36 Monate)
- SPK-KI Server:
  - 2x Nvidia Tesla V100 24GB (VGPU)
  - 36x Intel XEON 2.7 Ghz
  - 192GB RAM
- Freie Bereitstellung von SBB Daten, Technologien und Anwendungen: <https://github.com/qurator-spk>
- Mehr zur SBB in QURATOR: <https://qurator.ai/partner/staatsbibliothek-zu-berlin/>
- SBB Blogserie „Künstliche Intelligenz“: <https://blog.sbb.berlin/tag/wissenschaftsjahr-2019/>



# Metadatenanalyse

- Analyse von Metadaten (METS/MODS) und bestehenden Volltexten (ALTO) zur Gewinnung von Informationen über Merkmale und deren Verteilung



WC-Analyse

Mean squared error: 0.0040  
R<sup>2</sup> explained variance score: 0.50

	y_test	y_pred
PPN736019553	0.440000	0.448807
PPN622773135	0.693333	0.727628
PPN821772082	0.850000	0.673339
PPN717743837	0.472500	0.525989
PPN777540037	0.627500	0.650604
PPN666470626	0.665000	0.696438
PPN679959122	0.751111	0.661616
PPN735327963	0.557500	0.527351
PPN655207724	0.555556	0.637708
PPN816552711	0.521786	0.520847
PPN799319708	0.616667	0.617835
PPN755765214	0.720000	0.656037
PPN749791160	0.628333	0.692473
PPN746510098	0.553333	0.576473
PPN728337703	0.757143	0.705831

```

0.5630313527894495 originInfo-publication@_dateIssued
0.06318379892891454 classification-ZVDD_Ostasiatica
0.03823449159666124 originInfo-publication@_place_placeTerm_London
0.029608480124117924 language_languageTerm_ger
0.02942559461416441 originInfo-digitization@_dateCaptured
0.028823039309747406 subject-EC1418_genre_book
0.028199724369852127 classification-ZVDD_Rechtswissenschaft
0.023531983883285302 originInfo-publication@_place_placeTerm_Emden
0.023497157041718043 genre-aad_Liedersammlung
0.0229595813108022253 originInfo-publication@_publisher_Clark
0.02298749919793887 originInfo-publication@_place_placeTerm_Neudamm-Berlin
0.0157073500960244 file_count
0.013657140133976932 word_count
0.010590282448155778 classification-ZVDD_Kunst
0.009108230229550436 subject-EC1418_genre_journal
0.00834016966769741 classification-ZVDD_Musik
0.006893187886055325
0.005942084560844258
0.00568676772802392
0.004100967591525305
    
```

Feature Ranking

<https://github.com/qurator-spk/modstool>

	message
Changed w3cdtf encoding to iso8601	101757
Fixed eventType for a created origin	43254
Fixed eventType for electronic ed.	38851
Forced single instance of {http://www.loc.gov/mods/v3}dateCreated	36854
Fixed eventType for an issued origin	14300
Not a iso8601 date: "o.D."	2466
Forced single instance of {http://www.loc.gov/mods/v3}dateIssued	2105
Filtered {http://www.loc.gov/mods/v3}originInfo element (has no eventType)	1361
Forced single instance of {http://www.loc.gov/mods/v3}subject	762
Changed scriptTerm authority to lower case	437
Not a iso8601 date: "[18. Jh.]"	142
Not a iso8601 date: "1800 (1800c)"	124
Not a iso8601 date: "1780 (1780c)"	115
Not a iso8601 date: "[19. Jh.]"	86
Not a iso8601 date: "[15. Jh.]"	67

Metadaten-Validierung

# OCR Evaluation & Qualitätsverbesserung

- OCR Evaluation und Qualitätsverbesserung digitalisierter Dokumente durch bessere OCR und automatisierte OCR-Nachkorrektur

## Metrics

CER: 0.0135

WER: 0.07

## Character differences

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

## Word differences

Lorem ipsum dolor sit amet consetetur sadipscing elitr sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat sed diam voluptua At vero eos et accusam et justo duo dolores et ea rebum Stet clita kasd gubergren no sea takimata sanctus est Lorem ipsum dolor sit amet Lorem ipsum dolor sit amet consetetur sadipscing elitr sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat sed diam voluptua At vero eos et accusam et justo duo dolores et ea rebum Stet clita kasd gubergren no sea takimata sanctus est Lorem ipsum dolor sit amet

## OCR-Evaluation

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

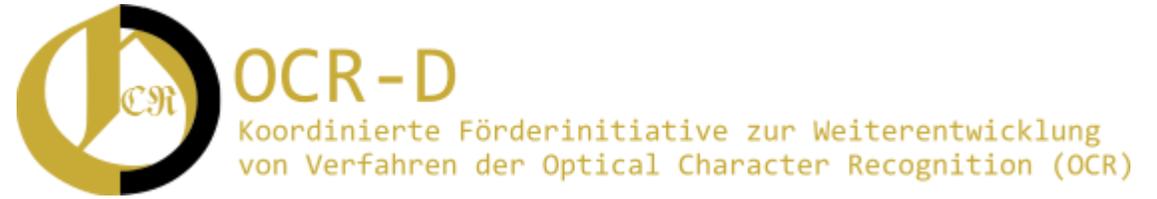
Lorem ipsum dolor sit amet consetetur sadipscing elitr sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat sed diam voluptua At vero eos et accusam et justo duo dolores et ea rebum Stet clita kasd gubergren no sea takimata sanctus est Lorem ipsum dolor sit amet Lorem ipsum dolor sit amet consetetur sadipscing elitr sed diam nonumy eimod tempor invidunt ut labore et dolore magna aliquyam erat sed diam voluptua At vero eos et accusam et justo duo dolores et ea rebum Stet clita kasd gubergren no sea takimata sanctus est Lorem ipsum dolor sit amet

$$\text{An examp} \left\{ \begin{array}{l|l|l} 1 & 0.8\% & I & 0.2\% & L & 0.0\% \\ 1 & 0.4\% & I & 0.5\% & L & 0.1\% \\ 1 & 0.2\% & I & 0.3\% & L & 0.2\% \end{array} \right\} e$$

Varianten-Voting  
in Calamari OCR

- <https://github.com/qurator-spk/dinglehopper>
- [https://github.com/qurator-spk/ocrd\\_calamari](https://github.com/qurator-spk/ocrd_calamari) (trainiert auf [GT4HistOCR](https://github.com/qurator-spk/GT4HistOCR))

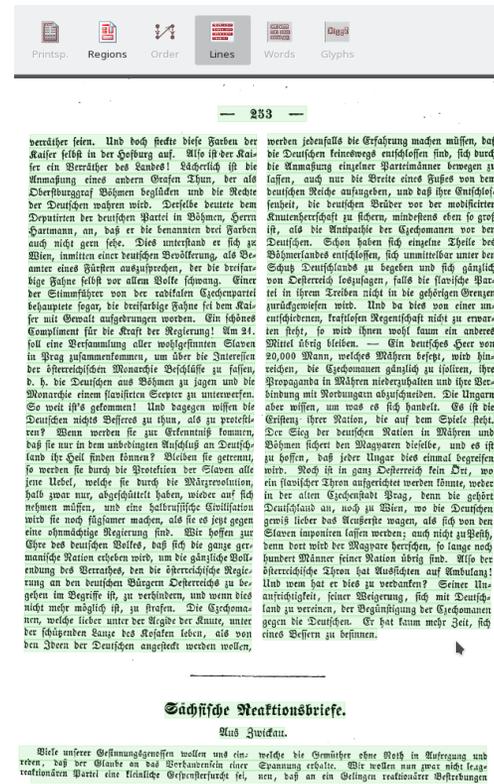
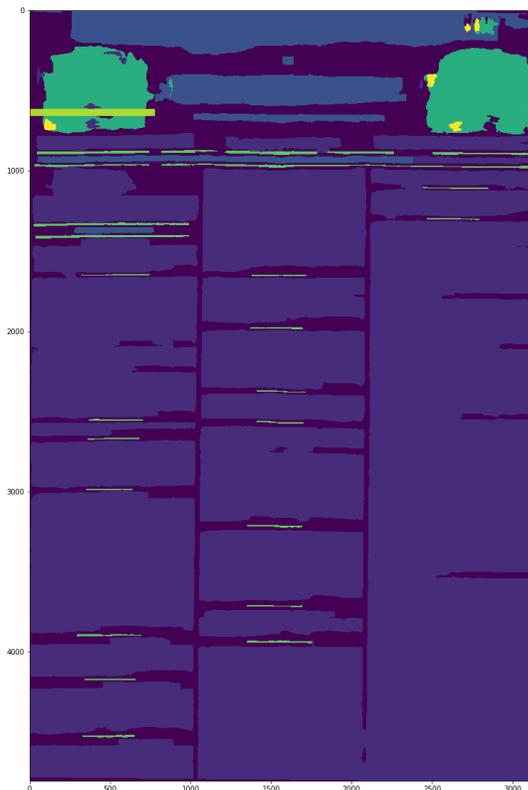
# OCR Exkurs: OCR-D



- DFG-gefördertes Koordinierungsprojekt + 8 verteilte Forschungsprojekte zu OCR für historische Drucke
- Webseite: <http://ocr-d.de/>
- Module: <https://ocr-d.github.io/projects>
- Source Code: <https://github.com/OCR-D>
- Dokumentation: <https://ocr-d.github.io/>
- Ground Truth: <http://ocr-d.github.io/gt-repo>
- Chat: <https://gitter.im/OCR-D/Lobby>

# Layout- bzw. Strukturerkennung

- Erkennung und Klassifikation von Strukturmerkmalen:  
[https://github.com/qurator-spk/pixelwise\\_segmentation\\_SBB](https://github.com/qurator-spk/pixelwise_segmentation_SBB)



- Pixel-Labeling mit ResNet50/UNet (CNN) für aktuell 16 Objektklassen
  - Spalten, Absätze, Separatoren
  - Überschriften, Fußnoten, Marginalien
  - Tabellen, Grafiken
  - usw.
- Textzeilenextraktion für die OCR
- Erkennung der Lese- bzw. Artikelreihenfolge (Reading Order)

# Named Entity Recognition

- Erkennung und Klassifikation benannter Entitäten in digitalisierten Dokumenten mit BERT: [https://github.com/qurator-spk/sbb\\_ner](https://github.com/qurator-spk/sbb_ner)

Task: Named Entity Recognition

Model: DC-SBB + CONLL + GERMEVAL

PPN: 865468370

Ergebnis:

Karte. 2 kleine Seiten quer 8°. 1054 Berlin, 3. Juni 1904. 2 Seiten 8°. 1055 Ragaz, 4. August 1904. 4 Seiten 8°. 1050 Berlin, 23. September 1905. 2 Seiten 8°. 1057 Berlin, 17. März 1908. 2 Seiten 8°. 1053 Berlin, 13. März 1910. 2 Seiten 8°. 1059 Sammelmappe mit Dankschreiben für die Zusendung der von C. R. Lessing veranlasseten Prachtausgaben des Nathan und der Minna von Bamheilm und der Geschichte der Familie Lessing (1880, 1890, 1909). 2°. 1000 Milchstück, Gustav. Ansprache an C. R. Lessing, an dessen 81. Geburtstag gehalten in Meseberg den 11. September 1908. Abschrift. 7 Seiten 8°. 1061 Taufschein von Emma Lessing geb. von Gelbke, ausgestellt vom Pfarrer der Garnison kirche in Berlin Ziehe den 12. Mai 1851. 1 Seite 2°. 1062

Otto Lessing 32.1 Sterbe - Urkunde von Emma Lessing geb. von Gelbke. Auszug aus dem Sterbe - Haupt - Register des Standesamts zu Berlin I und II vom 25. November 1895. Berlin, 17. Juli 1897. 2 Exemplare. 2 Seiten 2°. i063 8. Otto Lessing, des Malers Carl Friedrich Lessing ältester Sohn, Bildhauer, geboren 24. Februar 1846 in Düsseldorf, gestorben 22. November 1912 in Berlin. Carl Gussow an Otto Lessing. München, 23. Januar 1901. 8 Seiten 8°. 1071 Freundschaftliches. Künstlerzweise. Adolf Hildebrand an Otto Lessing. München, 8. Mai 1909. 1 2 Seiten 8°. 1072 Auskunft über Otto Ludwigs Büste. Georg II. Herzog von Sachsen - Meiningen: eigenhändige Bemerkung über Otto Lessings Büste Beethovens im Meiningen Hoftheater. Meiningen 1909, aus den Bauakten des Herzöglichen Hoftheaters. 1 Seite 8°. 1073. Die Büste Beethovens s. uimst sich vortreffl. aus Ich wüßte nichts zu kritisieren Gg. Gedicht auf Otto Lessings Gruppe, Die Kreuzträger von Wilhelm Votz in Karlsruhe. 4 Seiten 8°. 1074 Bei der Ausstellung des Werks in Karlsruhe dem Künstler übergeben. E. Andre Mitglieder und Verwandte der Familie Lessing 1. Petrus Lessig, Küster und Leinweber zu Einsiedel im Erzgebirge, Anfang des 17. Jahrhunderts. Abschriften aus den Kirchenvisitationsakten des Königlich Sächsischen Haupt staatsarchivs zu Dresden von 1617 betreffend die Pfarre Einsiedel und den Kustos (Küster) Petrus Lessig

Loc. 2003. 3 Seiten 4°. 1077 Lessingsche Bücher - It. Handschriftensammlung. 21

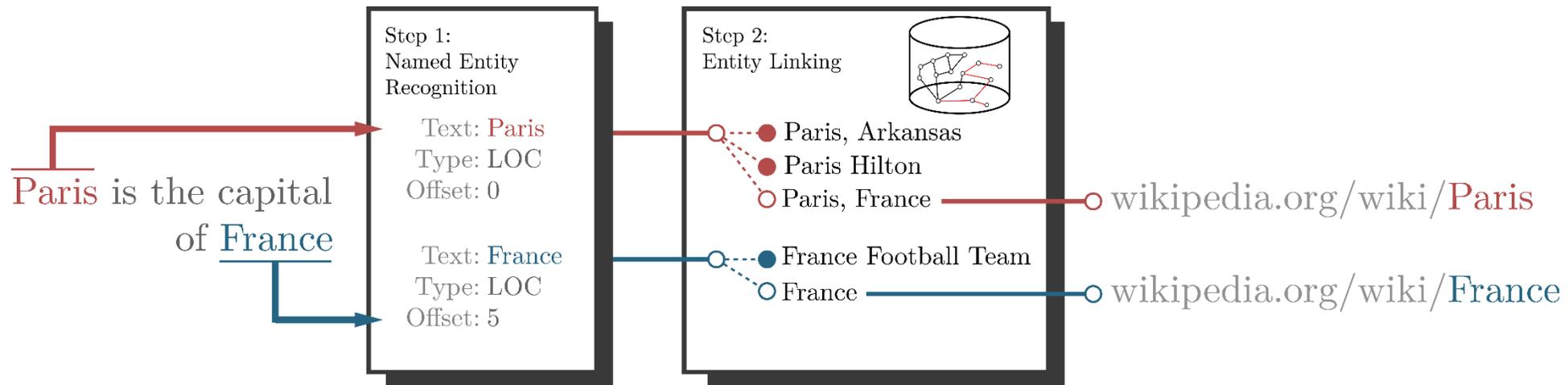
322 Die Kamener Leßing 2. Christian Gottlob Leßing, Bruder des Primarius Johann Gottfried Leßing, Bürgermeister von Kamen, geboren im April 1683 in Kamen, gestorben daselbst 10. Juli 1750. Registratur. Elstra, 15. April 1709. 4 Seiten 2°. 1073 Unterschriften unter Quittungen und andern Aktenstücken als Gerichtshalter zu Elstra, Administrator des Gestiftenamts, Prätor und Bürgermeister von Kamen. Elstra, 21. Juni 1712, 24. Oktober 1730; Kamen; Wal purgis 1715, 3. und 5. April 1732, Ostern 1741, 8. Dezember 1749; ein Blatt undatiert. 7 Blätter in 2°, 4°, quer 8°. 1079 -

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

5-fold cross validation on	pre-train	BERT multi-lingual-cased			(Riedl and Padó, 2018)	(Schweter and Baiter, 2019)
		precision	recall	F <sub>1</sub>	F <sub>1</sub>	F <sub>1</sub>
SBB	DC-SBB + GermEval + CoNLL	81.1 ±1.2	87.8 ±1.4	84.3 ±1.1	-	-
	DC-SBB + CoNLL	81.0 ±2.1	87.6 ±1.8	84.2 ±1.9	-	-
	DC-SBB + GermEval	80.6 ±1.8	87.4 ±1.3	83.8 ±1.2	-	-
	CoNLL	81.0 ±1.9	86.6 ±2.2	83.7 ±1.5	-	-
	GermEval	79.7 ±1.8	87.2 ±0.8	83.3 ±1.1	-	-
	GermEval + CoNLL	79.9 ±2.1	86.4 ±1.7	83.0 ±1.9	-	-
	DC-SBB	79.1 ±2.6	86.7 ±0.7	82.7 ±1.3	-	-
	none	79.1 ±3.6	85.0 ±1.1	81.9 ±2.2	-	-
ONB	Newspaper (1703-1875)	-	-	-	-	85.31
	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	-
LFT	none	78.0 ±2.4	84.1 ±1.9	80.9 ±2.0	73.31	-
	Newspaper (1888-1945)	-	-	-	-	77.51
	DC-SBB + CoNLL	70.0 ±2.6	81.0 ±0.7	75.1 ±1.5	-	-
	DC-SBB + GermEval	69.9 ±3.0	81.1 ±1.0	75.1 ±1.8	-	-
	DC-SBB	70.0 ±3.5	80.8 ±1.4	75.0 ±2.1	-	-
	DC-SBB + GermEval + CoNLL	69.8 ±3.0	80.8 ±0.9	74.9 ±2.0	-	-
	GermEval	68.9 ±2.7	79.3 ±1.4	73.7 ±1.9	74.33	-
	GermEval + CoNLL	69.1 ±2.6	78.8 ±1.3	73.6 ±1.5	-	-
none	68.8 ±3.4	79.2 ±1.5	73.6 ±2.2	69.62	-	
CoNLL	68.4 ±3.1	79.1 ±1.3	73.3 ±2.1	72.9	-	

# Named Entity Disambiguation & Linking

- Disambiguierung und Verlinkung benannter Entitäten mit einer Knowledge Base (Wikidata, GND)
- Erster Ansatz basierend auf Embeddings (Fasttext & Flair)



# Daten & Modelle

- <https://lab.sbb.berlin/>

The screenshot shows the Zenodo interface. At the top, there's a search bar and navigation links for 'Upload' and 'Communities'. Below that, the title 'Data and Demos of the Staatsbibliothek zu Berlin - Berlin State Library' is displayed. Under 'Recent uploads', two items are listed:

- June 26, 2019 (1.0)** | Dataset | Open Access | View  
**OCR fulltexts of the Digital Collections of the Berlin State Library (DC-SBB)**  
Labusch, Kai; Zellhöfer, David.  
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 (Luis/Baldwin 2012).  
Uploaded on June 26, 2019
- May 10, 2019 (v1)** | Technical note | Open Access | View  
**What is a PPN and Why is it Helpful?**  
Zellhöfer, David.  
Technical details regarding the handling of PPN (Pica production numbers) in the scope of the Berlin State Library. PPN can be used to download various digitized manifestations of the library's media such as full-page digitizations, OCR data, or derivatives such as JPEGs. PPN also allow a linkage  
Uploaded on May 10, 2019

A 'New upload' button is visible on the right side of the page.

- <https://zenodo.org/communities/stabi/>

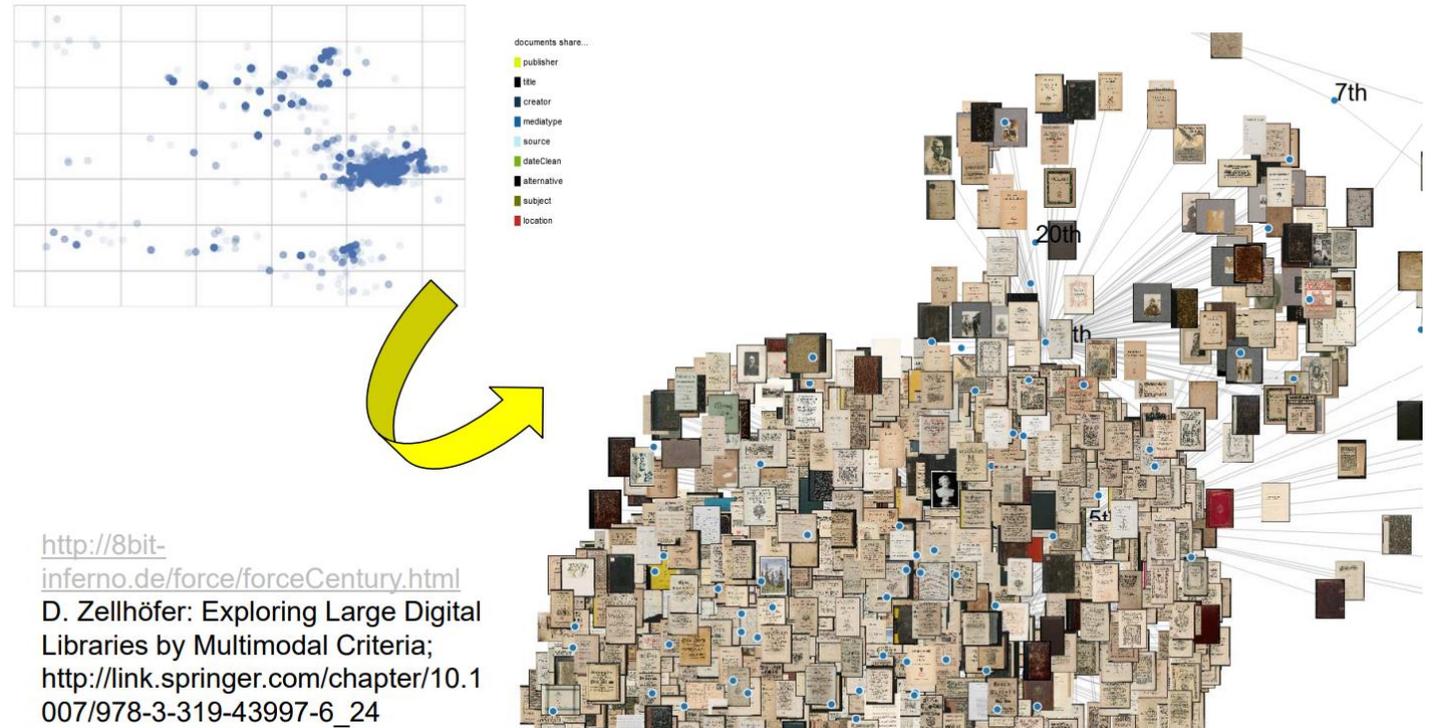
The screenshot shows the SBB-Lab website. The header includes the logo 'Staatsbibliothek zu Berlin Preussischer Kulturbesitz' and 'SBB-Lab'. Navigation tabs are 'HOME', 'DATEN', 'DEMOS', and 'KONTAKT'. The main content area is divided into several sections:

- Willkommen im SBB-Lab**: A welcome message stating that the digital revolution brings new knowledge and data interfaces, and that the lab provides selected datasets in an experimental environment. It also mentions a blog for more information.
- Daten**: A section titled 'Downloads und Schnittstellen' (Downloads and Interfaces) with the text: 'Entdecken Sie von uns zusammengewürfelte Inhalte verschaffen können'. Below this is a 'Bibliographische Daten StabiKat' section, which is a catalog of books, journals, and electronic media from 1500 to 1945, totaling 14 million items.
- Demos**: A section titled 'Projekte und Prototypen' (Projects and Prototypes) with the text: 'Experimentelle Anwendungen, Projekte und Prototypen lernen Sie im Demobereich unseres SBB-Lab kennen. Da einige Inhalte allerdings nicht von uns erstellt wurden, leiten wir Sie teilweise auf externe Seiten - vielleicht ja schon bald auch auf Ihre!'
- Altpapier**: A section for 'Altpapier' (Old Paper) with the text: 'Die App Altpapier bringt einige der spannendsten und merkwürdigsten Zeitungsmeldungen des frühen 20. Jahrhunderts auf die Smartphones.' Below this is a 'Berliner Schlagzeilen' section, which is a bot that tweets daily headlines from the Berlin Schlagzeilen newspaper.

At the bottom right, there is a language selector set to 'Deutsch'.

# Ausblick

- Bildähnlichkeitssuche unter Verwendung von VGG16 und Re-training mit ImageNet
- Geolokalisierung durch Kombination von semantischen mit topographischen Merkmalen
- Demonstrator basierend auf digitalisierten Sammlungen (SBB Lab)



# Danke für die Aufmerksamkeit! Fragen?

Clemens Neudecker ([@cneudecker](#))

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Fachtagung Netzwerk maschinelle Verfahren in der Erschließung

10.-11. Oktober 2019, DNB, Frankfurt a.M.