

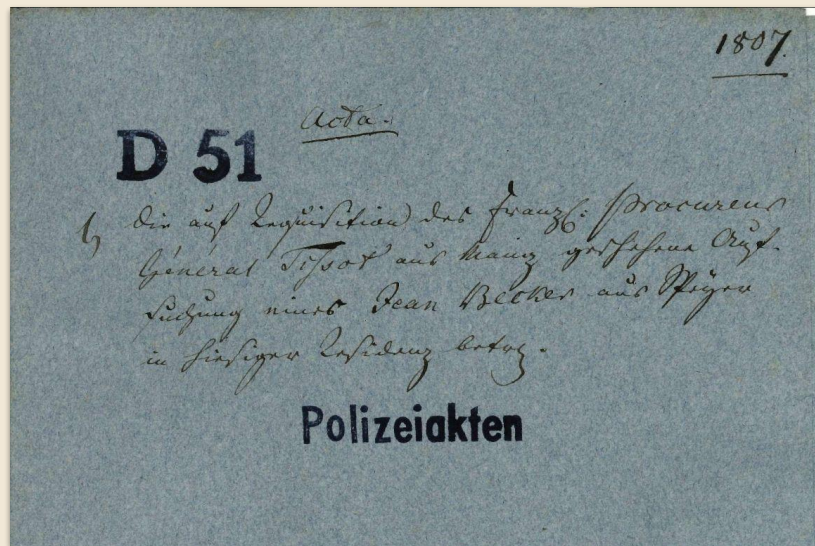


Saubere, kontrollierbare und zuverlässige Daten- und Informationsextraktion für Archive

Beitrag zum 29. Archivwissenschaftlichen Kolloquium der Archivschule Marburg,

Michaela Prien, **Transkribus**, READ-COOP SCE

Die Herausforderung



Quelle: Landesarchiv Baden-Württemberg, Abt. Staatsarchiv Ludwigsburg D 51 Bu 17 Bild 1; Permalink <https://www.landesarchiv-bw.de/de/ku/01029929015>; Rechtshinweis: Public Domain Mark 1.0 (<https://creativecommons.org/licenses/by/4.0/>)

1807

Acta

die auf Requisition des **Franzl. Procureur General** Tissot aus Mainz geschehene Auf-
suchung eines Jean Becker aus Speyer
in hiesiger Residenz **betrl.**

Polizeiakte

[Transkribus, Text Titan I, 16.05.2025]

D 51

Acta.

Betr. Auf Ersuchen des französischen Procureur
Général **Toffot aus Nancy getroffene Verfügung**
betreffend Jean Becker **aus Bingen**
in hiesiger **Haftanstalt** betr.

Polizeiakten

(Rechts oben handschriftlich: 1807)

[ChatGPT 4o, 16.05.2025]

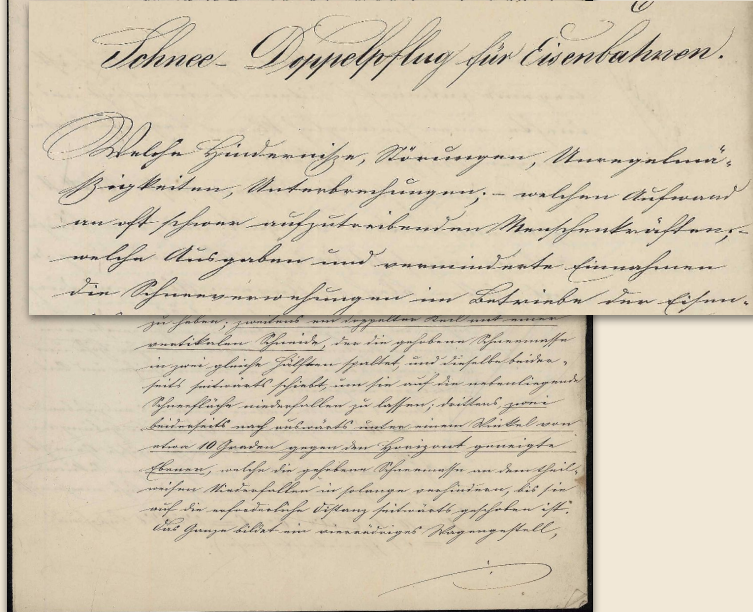
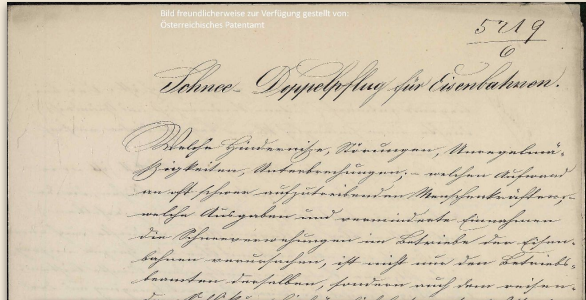
1807

D 51 Acta

betr. **Aufsichtssache** des franz[ösischen] Procureurs
General Tissot und Banquier Procureur Vogt.
Aufsatz nicht Dean Becker **aus Pfizen**
in hiesiger **Polizeiakten** betr[effend].

Polizeiakten

[Gemini Advanced, 2.0 Flash, 16.05.2025]



Schnee-Doppelpflug für Eisenbahnen.

Welche Hindernisse, Störungen, Unregelmäßigkeiten, Unterbrechungen; – welchen Aufwand an oft schwer aufzutreibenden Menschenkräften, – welche Ausgaben und verminderte Einnahmen die Schneeeverwehungen im Betriebe der Eisen-

[Transkribus, Text Titan I, 15.05.2025]

Schnee-Doppelpflug für Eisenbahnen.

Obgleich Spureinreißer, Räumungswagen, Rotationsräumer, Schneepflüge, Schneeanwerfmaschinen u. mehrere derartige um oft schwere mitführbare oder Menschenkraft erfordernde andere Hilfsmittel zur Freimachung oder Freihaltung der Bahnstrecken gegen den Schnee oder

[Claude 3.7, 15.05.2025]



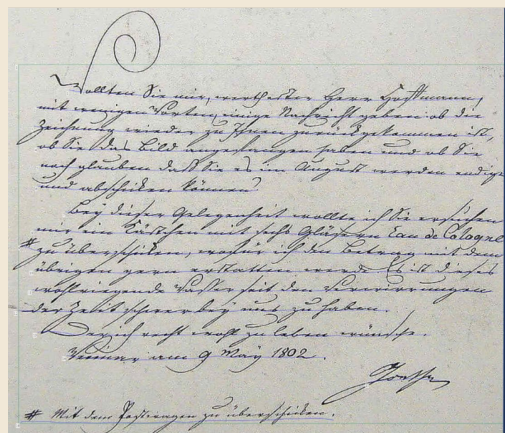
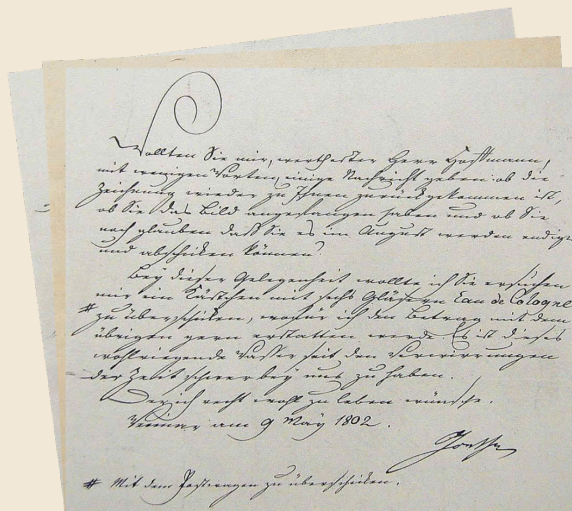
Wie jedes gute Werkzeug nimmt KI Arbeit ab –
Bedeutung und Kontext entstehen durch
fachliches Wissen.

- ① Aktueller Stand: Wofür KI im Archiv bereits eingesetzt wird
- ② Die nächste Generation: End-to-End-Modelle
- ③ Praxisbeispiele & Anwendungsszenarien
- ④ Ausblick: Integration großer Sprachmodelle (LLMs)



Aktueller Stand: Wofür KI im Archiv bereits eingesetzt wird

Texterkennung & Transkription



Wollten Sie mir, werthester Herr Hoffmann,
mit wenigen Worten, einige Nachricht geben: ob die
Zeichnung wieder zu Ihnen zurück gekommen ist,
ob Sie das Bild angefangen haben und ob Sie
noch glauben daß Sie es im August werden endigen
und abschicken können?

Bey dieser Gelegenheit wollte ich Sie ersuchen,
mir ein Kästchen mit sechs Gläsern Eau de Cologne
zu übersenden, wofür ich den Betrag mit dem
übrigen gern erstatten werde. Es ist dieses
wohlriechende Wasser seit den Verwirrungen
der Zeit schwer bey uns zu haben.
Der ich recht wohl zu leben wünsche.

Weimar am 9. May 1802.

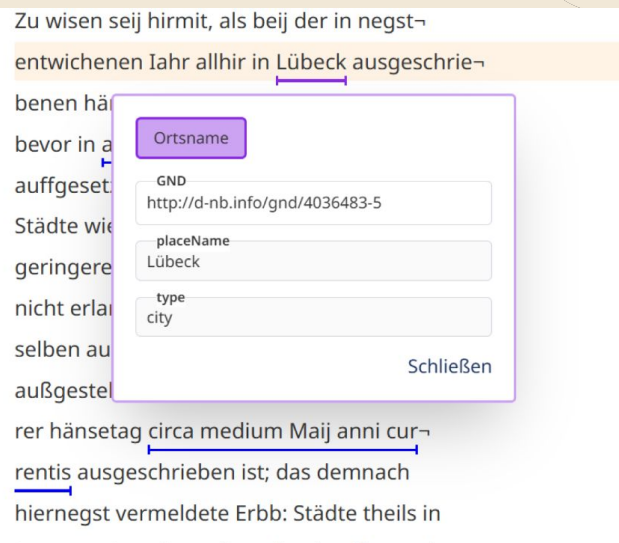
Goethe

Mit dem Postwagen zu übersenden.

Strukturen erkennen (Formulare, Tabellen, Marginalien)

Kyoto Academy	Factorville, Pa.	5000	Gymnasium	1/6/11	Low
Keller College	Philadelphia, Pa.		General	1/1/11	Low
Kemper College	P Gambier, O.		Bildung	2/18/11	Has done share, 2/20/11 ; D
Kansas Christian College	Lincoln, Kan.	10 000	Endowment	2/18/11	Low
Kentucky State Univ. of (1) C	Louisville, Ky.	600	Teacher's salary	2/4/11	Not sufficient to be paid. 2/20/11
Knox Institute & Ind. School (1) C	Altoona, Pa.				
Kols Normal & Training Inst.	Kols				
Kilder Institute	Kilder				
Kansas Univ. of	Laurin				
Kansas Wesleyan University (1) F	Salina				
Kentia College (1)	Kentia				
Kingfisher College (1)	Kingfisher				
Knox Institute & Ind. School (1) F	Altoona				
Knoxville College	P Knoxville				
King College (1)	Smith				

Haupt-Grundbuchheft (Offentjahrgang)		1904	Blatt-Nr.		625		
Vor- und Zuname		Name: Johann Hüpfauß Hüpfauß					
Geburts-	Ort	Innsbrück	Ortsgemeinde	Innsbrück	Geburts-jahr	1883	
	Bezirk	Innsbrück		Bezirk	Innsbrück	Religion	kathol.
	Comitat	%		Comitat	%	Kunst, Gewerbe, sonstiger Lebensberuf	Lehrer
	Land	Tirol		Land	Tirol		
nach der Losreihe auf drei Jahre in der Reserve und zwei Jahre in der Landwehr, zum 3. April d. Tirol. Karl. Lager							



Ortsname

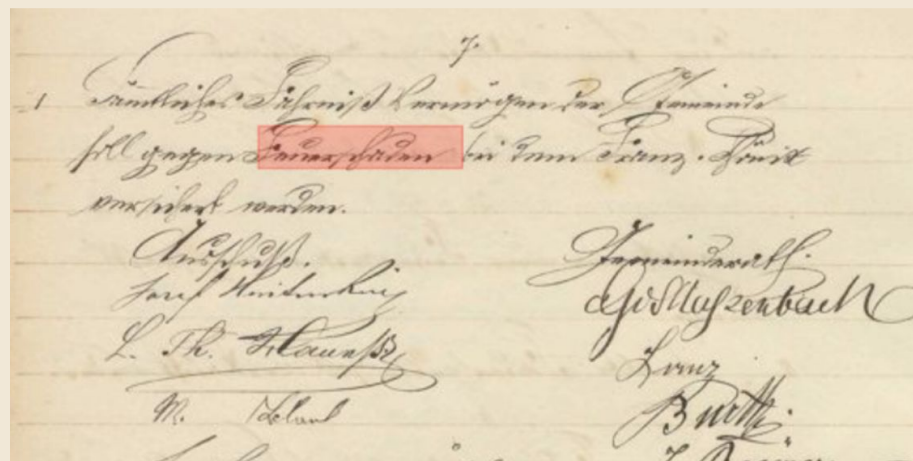
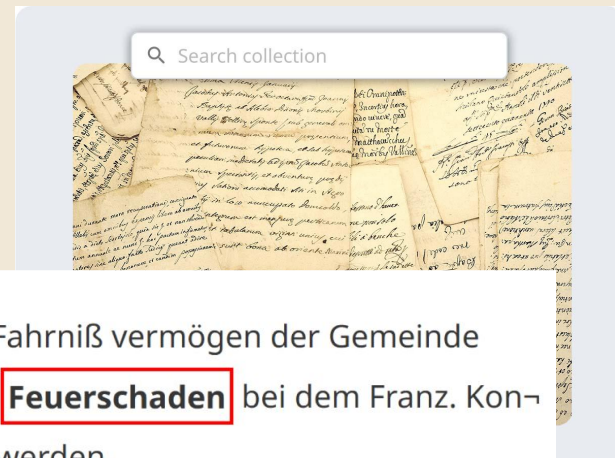
GND
http://d-nb.info/gnd/4036483-5

placeName
Lübeck

type
city

Schließen

Durchsuchbarkeit & Recherchehilfe



ämtliches Fahrniß vermögen der Gemeinde
soll gegen **Feuerschaden** bei dem Franz. Kon-
versichert werden.
Ausschuß.
Gemeinderath
Forch Unterkie-
Gassenbach



Die nächste Generation: End-to-End-Modelle

Die Herausforderung:

Wie kommen wir möglichst einfach zu sauberen und zuverlässigen Daten und Informationen?



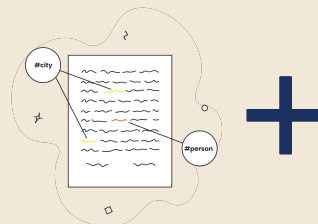
Die Lösung:

Saubere, kontrollierbare und zuverlässige Daten- und Informationsextraktion mit End-to-End-Modellen



Was bedeutet End-to-End?

Bisheriger sequentieller Workflow:



Hochladen

Layouterkennung:

- Zeilen
- Felder
- Tabellen

Texterkennung

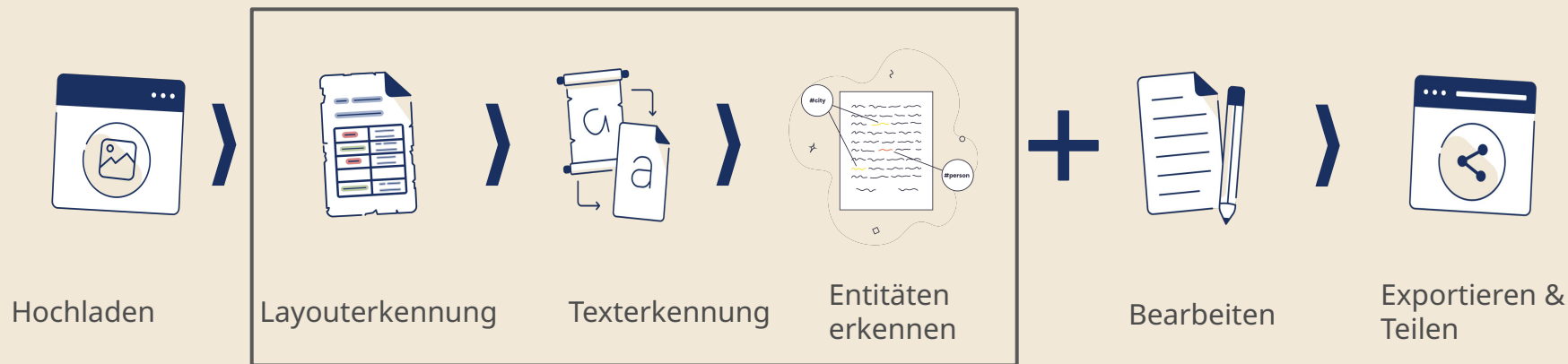
Entitäten
markieren

Bearbeiten

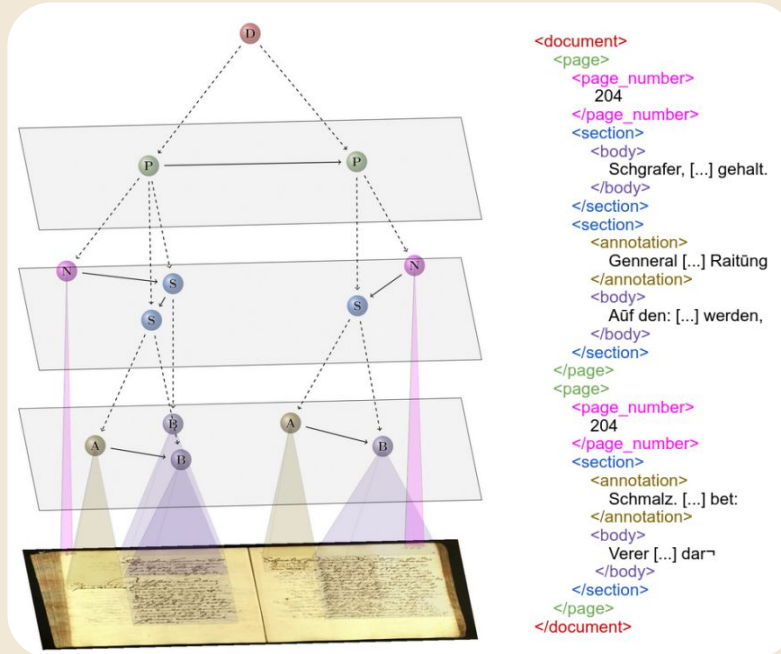
Exportieren &
Teilen

Was bedeutet End-to-End?

End-to-End Workflow:



Wie End-to-End Modelle arbeiten



- Betrachten das gesamte Bild in einem Schritt
- Erkennen Text, Layout und hierarchische Strukturen gleichzeitig
- Liefern strukturierte, kontrollierbare Informationen – ohne Zwischenschritte

End-to-End heißt: Ein Modell, ein Durchgang, ein Ergebnis – direkt vom Bild zur Information.



Praxisbeispiele & Anwendungsszenarien

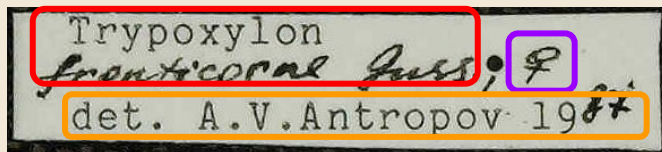
Beispiel: Etiketten von Sammlungsobjekten



- Zeichenfehlerquote (CER): 27 % → 8 % (bei gemischten Schriften)

Beispiel: Etiketten von Sammlungsobjekten

© MfN



- Region 1
- 1 Trypoxylon
 - 2 fronticorne Guss., ♀
 - 3 det. A. V. Antropov 1987



Specimen	Sex	Determination
Trypoxylon fronticorne Guss.	♀	det. A.V. Antropov 1987

- Zeichenfehlerquote (CER): 27 % → 8 % (bei gemischten Schriften)
- Gute Datengrundlage für Entitätenerkennung

Beispiel: Tabellarische Daten

Kopfzeile/header

Patienteninformation/
patient items

NATIONAL HEALTH MANAGEMENT INFORMATION SYSTEM

1. STATE: KAD 2. LGA: Zaria 3. WARD: Saku

6. S/N	7. Date (DD/MM/YYYY)	8. Name of Patient/Client	9. Patient's Care Number	10. Sex (Tick as appropriate)	11. Age (In appropriate above age range column)					12. Type of Admission (Tick as appropriate)	13. Nutritional Assessment		14. Presenting Complaint(s)	15. Type of Laboratory Investigation Done	16. Diagnosis	17. Drugs Given	
					2-28 days	29 days - 12 months	12 months - 59 months	6-9 years	10-19 years		20-59 years	60 years and above					Weight (kg)
1	9/11/14	SUREN ALI		✓						24	✓	36	140	Backpain, fever	RBT	malaria	PCM, AS
2	11/11/14	KABIR MAGEZ		✓						36	✓	41	140	Severe	RBT	malaria	AS
3	11/11/14	BELLO BASHIR		✓						28	✓	19	140	Chills, fever	RBT, Widal	typhoid, malaria	PCM, AS
4	11/11/14	SALIMU JABAR		✓						24	✓	91	140	Febrile	Widal	malaria	AS
5	11/11/14	SALIMU SAKU		✓						15	✓	36	140	Accident	RBT	typhoid, malaria	AS
6	11/11/14	JABAR SAKU		✓						12	✓	42	140	Severe	RBT	malaria	AS
7	11/11/14	SALIMU SAKU		✓						14	✓	42	140	Fever	RBT, Widal	malaria	AS
8	11/11/14	BELLO MAGEZ		✓						23	✓	21	140	Severe	RBT	malaria	AS
9	11/11/14	BELLO MAGEZ		✓						36	✓	14	140	Backpain, fever	RBT	malaria	AS
10	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
11	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
12	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
13	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
14	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
15	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
16	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
17	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
18	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
19	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS
20	11/11/14	BELLO MAGEZ		✓						24	✓	42	140	Severe	RBT	malaria	AS

Page Total: 61

20. NAME OF REPORTING OFFICER: Sakha Ali 21. DESIGNATION: Nurse



```

1 {
2   "header": {
3     "1_State": "kad",
4     "2_LGA": "zaria",
5     "3_Ward": "saku"
6   },
7   "items": [
8     {
9       "13_Weight": "36",
10      "14_Complaints": "backpain, fever",
11      "16_Diagnosis": "malaria"
12    },
13    {
14      "13_Weight": "41kg",
15      "14_Complaints": "fever",
16      "16_Diagnosis": "malaria"
17    },
18    {
19      "13_Weight": "19",
20      "14_Complaints": "vomiting, fever",
21      "16_Diagnosis": "malaria, typhoid"
22    },
23    {
24      "13_Weight": "91",
25      "14_Complaints": "typhoid fever",
26      "16_Diagnosis": "malaria"
27    },
28    {
29      "13_Weight": "36",
30      "14_Complaints": "accident",
31      "16_Diagnosis": "typhoid"
32    },
33  ]
34 }

```

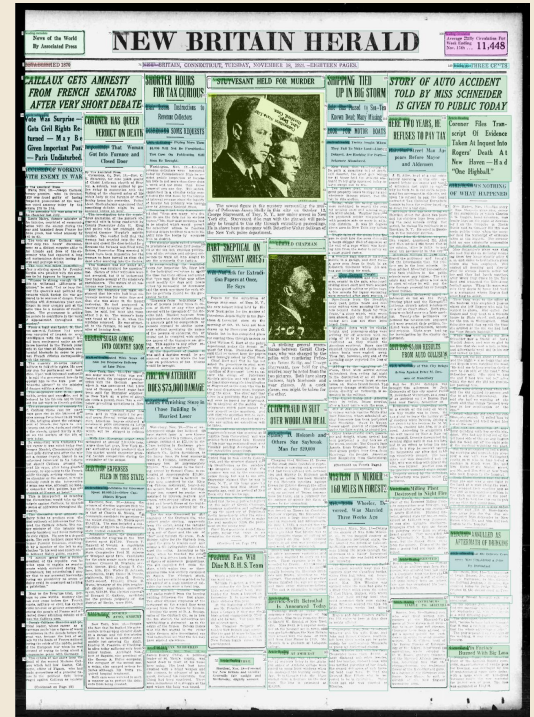
- Zeichenfehlerquote (CER): 4%
- Wortfehlerquote (WER): 2%

Beispiel: Zeitungen

Oberste Ebene

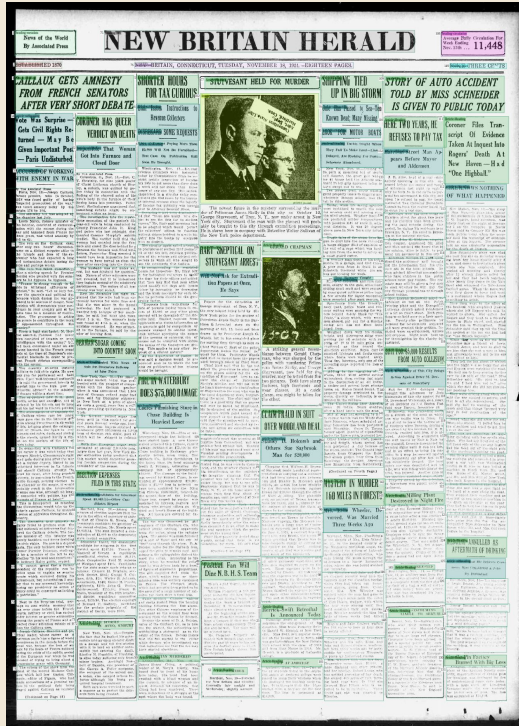
Mittlere Ebene

Unterste Ebene



Beispiel: Zeitungen - Strukturierte Ausgabe

Input



Output

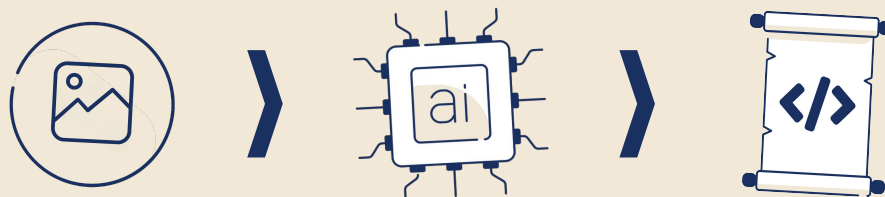
```

1 <Page>
2 <headings>
3   <heading-top>
4     <heading-metadata>
5       News of the World
6       By Associated Press
7     </heading-metadata>
8     <heading-title>
9       NEW BRITAIN HERALD
10    </heading-title>
11    <heading-circulation>
12      Average Daily Circulation For
13      Week Ending
14      <tag_date>Nov. 15th...</tag_date>
15      11,448
16    </heading-circulation>
17  </heading-top>
18  <heading-bottom>
19    <heading-established>
20      ESTABLISHED <tag_date>1870</tag_date>
21    </heading-established>
22    <heading-date>
23      NEW BRITAIN, CONNECTICUT, TUESDAY, <tag_date>NOVEMBER 18, 1924</tag_date>...EIGHTEEN PAGES.
24    </heading-date>
25    <heading-price>
26      PRICE THREE CENTS
27    </heading-price>
28  </heading-bottom>
29 </headings>
30 <body>
31   <article>
32     <article-heading>
33       <tag_name>CAILLAUX</tag_name> GETS AMNESTY
34       FROM FRENCH SENATORS
35     </article-heading>
36     <article-subheading>
37       AFTER VERY SHORT DEBATE
38     </article-subheading>
39     <article-text>
40       Vote was Surprise --
41       Gets Civil Rights Re-
42       turned -- May Be
43       Given Important Post
44       -- Paris Undisturbed
45     </article-text>
46     <article-subheading>
47       ACCUSED OF WORKING
48       WITH ENEMY IN WAR
49     </article-subheading>
50     <article-paragraph>
51       By the Associated Press.
52       Paris, <tag_date>Nov. 18</tag_date>...<tag_name>Joseph Caillaux</tag_name>,
53       former premier, who in October
54       <tag_date>1918</tag_date> was found guilty of having
55       "impeded prosecution of the war."
56       was voted amnesty today by the
57       senate 176 to 104.
58     </article-paragraph>
59   </article>
60 </body>
61 </Page>

```

Vorteile von End-to-End-Modellen

- Vereinen Transkription, Layout-Analyse, Extraktion strukturierter Informationen, Erkennung von benannten Entitäten in einem Modell
 - Keine Fehlerfortpflanzungen mehr → höhere Genauigkeit
- Ermöglichen hierarchische Entitäten-/Layout-Extraktion
 - Arbeit mit komplexen Leserichtungen, verschachtelten Strukturen etc. wird einfacher
- Zeitersparnis & Skalierbarkeit
 - Einfachere Wartung der Modelle möglich → weniger Overhead
 - schnellere Verarbeitung großer Mengen, z. B. bei Massenakten oder Zeitungsjahrgängen
- Sind kontrollierter im Verhalten als große Sprachmodelle





Anwendung von End-to-End-Modellen



Saubere, kontrollierbare und zuverlässige
Daten und Informationen



Ausblick: Integration großer Sprachmodelle (LLMs)

End-to-End als Schlüssel für:

- die Weiterverarbeitung mit Hilfe großer Sprachmodelle (LLMs)
- das Trainieren großer Sprachmodelle
- die Unterstützung großer Sprachmodelle



LLM Dokumenten Analyse

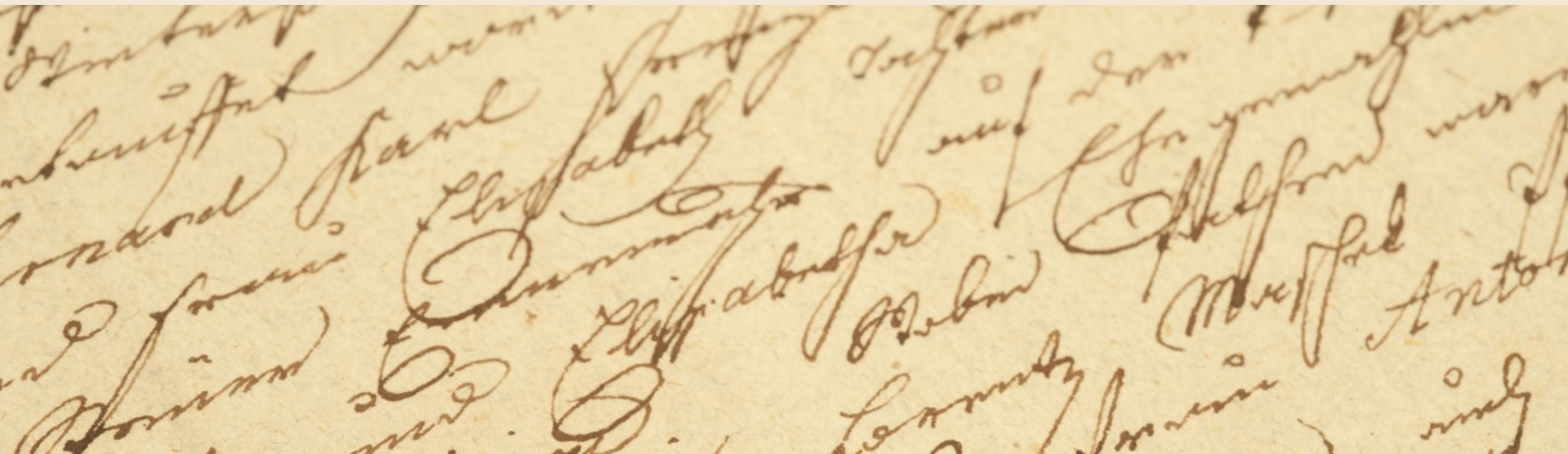


Korrektur

Übersetzung

Zusammenfassung





Geschichte entschlüsseln mit Transkribus

m.prien@transkribus.org

<https://www.transkribus.org/>