

Transkribus & eScriptorium

Zwei Plattformen zur Transkription und automatisierter Erkennung von
historischen Textquellen

Larissa Will und Jan Kamlah (Universität Mannheim)

Dr. Stefan Büdenbender (Hochschule Darmstadt)

2. Vortrag und Hands-On zur 2. Staffel der Vortragsreihe

Von Büchern zu Bytes

22. November 2024, 10-11:30 Uhr online



Im Namen des 4Memory-Konsortiums

Transkribus & eScriptorium

Zwei Plattformen zur Transkription und automatisierten Erkennung
von historischen Textquellen



Transkribus[®]

337 812 074 **M** versichert blieben, davon 552 246
1 847 622 742 **M** bei den 35 Se im
b. Der gesammte Zuwachs im Jahre 1877 stellt sich
tionen mit 110 873 820 **M** d. h. 2,14 % der
Zahl der Versicherten und 4,98 % der Wachen,
summe. Die Berücksichtigung der Unfälle, welche die
Liquidation einiger Gesellschaften im Jahre 1877 zur Folge gehabt
hat vermindert sich der Zuwachs auf 8542 (1,16%) der Versicherten



Larissa Will, Jan Kamlah und Stefan Büdenbender

Transkription/OCR/HTR: kurze Einführung und Verortung

In welchem Kontext finden Transkription/OCR/HTR statt?

- Volltextdigitalisierung von Kulturgütern
- Sehr unterschiedliche Nutzungsszenarien in Bezug auf:
 - Anwender*innen und Ressourcen: Einzelforschende, (Citizen-Science-)Projekte, Gedächtniseinrichtungen, Digitalisierungsinitiativen...
 - Ausgangsmaterial: Umfang, Materialität, Schriftart, Qualität...
 - Fachdisziplinen und Forschungsfragen: historisch arbeitende Disziplinen, Sprach- und Literaturwissenschaft, Corpuslinguistik,
 - Intendierter Nutzen/Zielgruppe: menschliche Leser*innen, Forschende, Information Retrieval...
 - Qualitätsanspruch: „schmutziges“ OCR/HTR, automatische/manuelle Korrekturen, Anreicherung mit Metadaten...
 - Formate: Plain text, PDF (Bild+Text), XML mit Annotationen/Metadaten
 - Endprodukt: (lokale) Dateisammlung, Korpus, digitale Edition

Wo ist die (automatisierte) Transkription anzusiedeln?

- I. Digitale, **maschinenlesbare Metadaten** ohne Digitalisate
- II. Digitale Repräsentation von analogen Daten – **Bild-Digitalisate**
- III. **Volltexte unstrukturiert** bzw. nicht standardisierte Strukturierung
 - A. Roh-OCR
 - B. optimierte OCR (> 99 % Zeichengenauigkeit)
 - C. Manuelle Erfassung (> 99,97 % Zeichengenauigkeit)
 - D. Volltexte (per OCR/OLR und/oder manuell erzeugt)
- IV. **Volltexte mit standardisierter Strukturierung** (maschinell weiterverarbeitbar)
- V. **Volltexte angereichert, z.B. mit Entitäten und komplexen Strukturmerkmalen**
 - A. angereichert mit Normdaten (z.B. GND, VIAF, TNG, ...)
 - B. Annotation
 - C. Kollation
 - D. Digitale Edition

Tools wie
Transkribus
und
eScriptorium
können weite
Bereiche
davon
innerhalb der
Plattform
abdecken.

Mehrwert hoch angereicherter Daten: Beispiel Digitale Edition

WERKAUSWAHLDER PUPPENSPIELERDOKUMENTARISCHE ANSICHT ZU PUP_pH1

BETA 3.0MITWIRKENDEIMPRESSUMZUR PROJEKTSEITE

ASdÜBER DIE EDITION


ERWEITERTE SUCHE

PUP_pH1_0001 / 2

ARBEITSSTADIUM: Letztgültiger Textzustand

Annotationen (TRANSKRIPTBEZOGEN)

Messias



Grundzustand

1. Hand Arthur Schnitzler, Tinte

| Ebene | Annotation | Wert | Textstufe | Position |
|-------------|-----------------------------|----------------|-------------|-----------|
| Handschrift | Schriftsystem | lat. Kurrent | | [Messias] |
| Handschrift | Einfache Unterstreichung | Einfache Linie | Textstufe 1 | [Messias] |
| Handschrift | Sofortänderung | | | [Messias] |
| Handschrift | Intendierte Unterstreichung | Einfache Linie | Textstufe 1 | [Messias] |
| Handschrift | Sofortänderung | | | [Messias] |

Bearbeitungszustand

1. Hand Arthur Schnitzler, Bleistift

1 Cavalier verarmt; Gläubiger statten ihn

2 zur Werbung aus; Geliebte reist nach.

3 Sein Abenteuer, ein Mädel wird ihm des Späßes

4 halber von seinen Freunden zugeführt; er nimt's

5 ernst.

6 Wittwer muß er – für sein Werk – werden –

7 heiratet eine todtkranke, verliebt sich in sie.

8 Ueb Pers. Kein Kind. Du bist ja verheiratet; gib dich ein-

9 mal deinem Maß hin – Sie erstarrt, spuckt ihn

10 an Wahrheit sie gab sich einem Knecht hin; muß

11 sie's dem Maß gestehen? – es war ja nichts. –

12 Messias würden deine Frau u ich erzeugen –

13 gib sie mir! – Kind Cretin.

14 Gefordert, weil er seine Frau verführte. Un-

15 wahr. Sie kömt. Nimm mich u fliehen wir!

16 Er bleibt u wird erschossen.

17 Componist, Erfolg der Oper, kömt in die Loge

18 seiner Frau, findet sie mit ihrem Geliebten!

19 Ehebrecherin vor der Ehe. Verdacht des Maßs.

20 Berührt sie nicht. Kind zu früh. – Erwürgt es. –

21 Mädchen, die sich einbildet, er starb um

22 ihretwillen u starb nur um eine Dirne! –

23 Faustine sucht einen leidenschaftlichen

24 Mann; wartet auf einen, der seine Geliebte

25 tödtet vor Eifersucht.

26 Von seiner Tänzerin möcht er los. – Theater breiß –

27 er läuft sie retten u kömt um. –

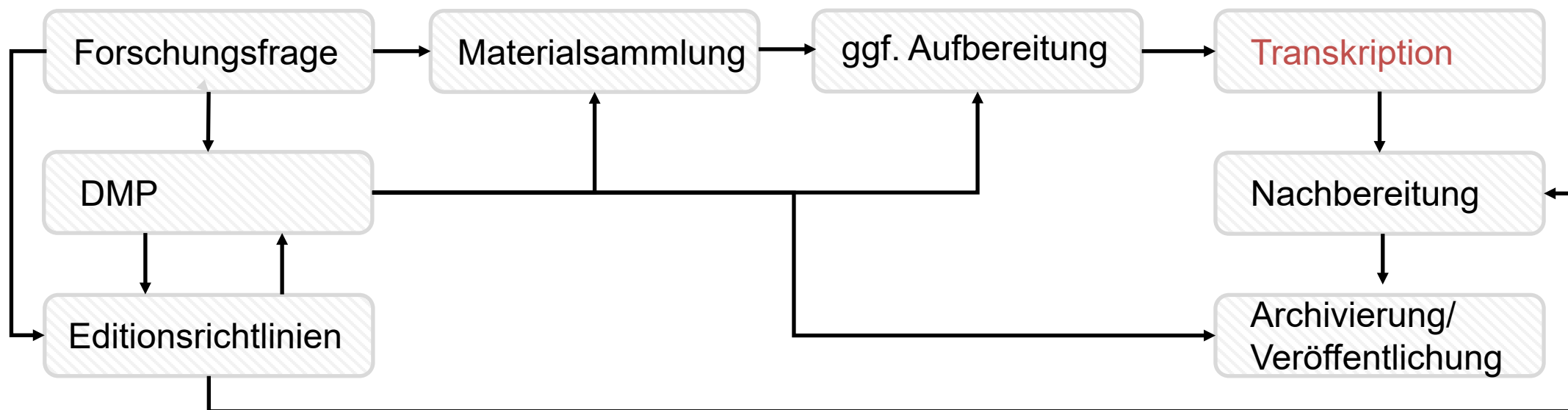
BETA VERSION 3.0 STAND 31.01.2024

LETZTE INHALTLICHE AKTUALISIERUNG 16.01.2024

Larissa Will, Jan Kamlah, Stefan Büdenbender | 22.11.2024

<https://www.arthur-schnitzler.de>

Wie läuft ein Digitalisierungsprojekt mittels (automatisierter) Transkription ab?



Transkriptionsplattform Basics

Was ist Transkribus?

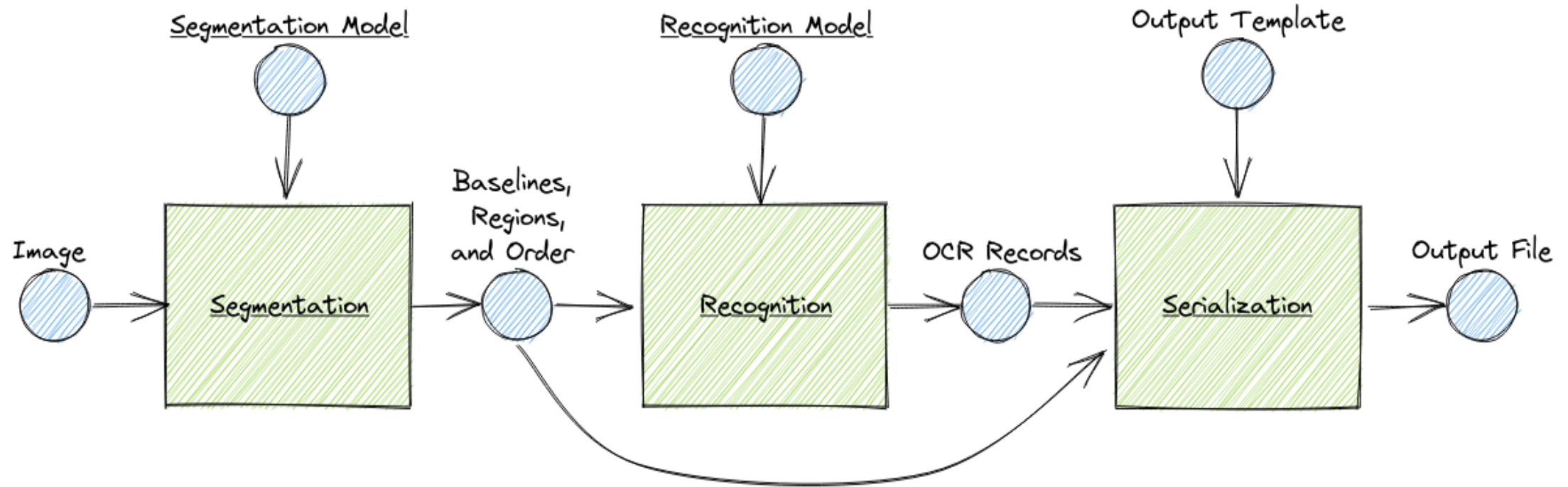


- Transkriptionsplattform (OCR-Engine: [PyLaia](#))
- entstand in den Projekten tranScriptorium und READ an der Universität Innsbruck
- Betrieb und Weiterentwicklung durch die READ-COOP
- Freemium Geschäftsmodell
- Für die Verarbeitung müssen Bilder auf die Server von Transkribus hochgeladen werden
 - die On-Prem (Vor-Ort, Verwendung eigener Infrastruktur) Lösung kann hier eine Abhilfe schaffen
- Handgeschriebene und gedruckte Texte manuell oder automatisiert segmentieren und transkribieren
- Modelle können nur auf der Transkribus-Instanz geteilt werden

Was ist eScriptorium?

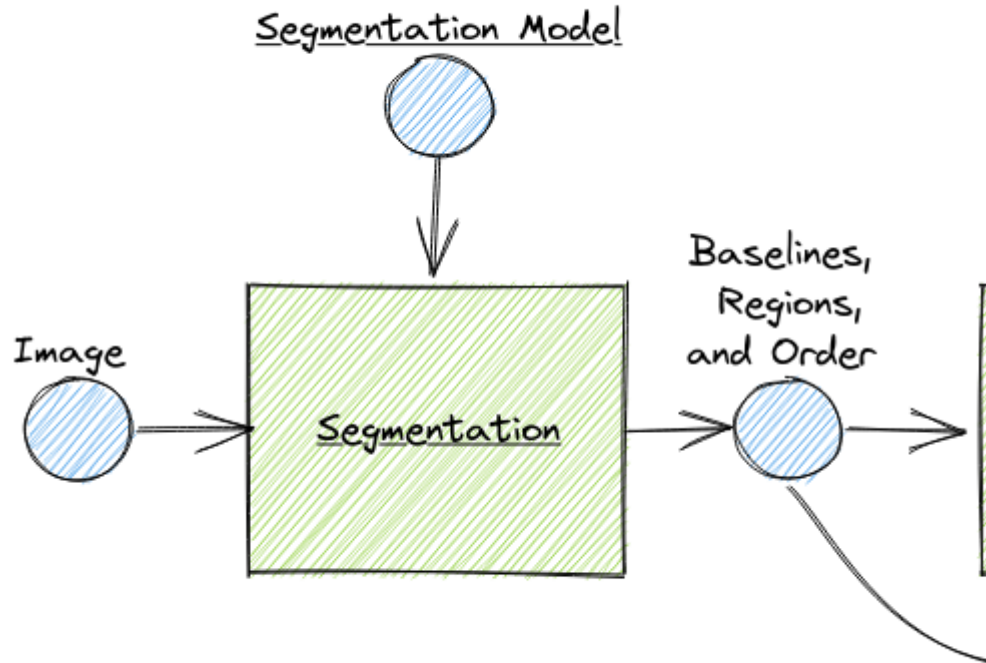
- Transkriptionsplattform (OCR-Engine: [kraken](#))
- entwickelt an der Université Paris PSL
- Kostenfrei und Open-Source
- Alternative zu Transkribus
- Handgeschriebene und gedruckte Texte manuell oder automatisiert segmentieren und transkribieren
- Einfache Weitergabe trainierter Modelle
- Jeder kann eigene Instanz installieren
 - Für Windows, MacOS und Linux





Bildcredit: <https://kraken.re/main/index.html>

Workflow-Schritt: Segmentierung



Unter **Layout-Segmentierung** versteht man den Prozess der Erkennung von Inhaltsblöcken aus dem Hintergrund eines digitalisierten Dokuments.

Unter **Regionsklassifizierung** hingegen versteht man die Einteilung der bei der Layout-Segmentierung gefundenen Blöcke in verschiedene Kategorien.

Bildcredit: <https://kraken.re/main/index.html>

Workflow-Schritt: Segmentierung



Workflow-Schritt: Segmentierung

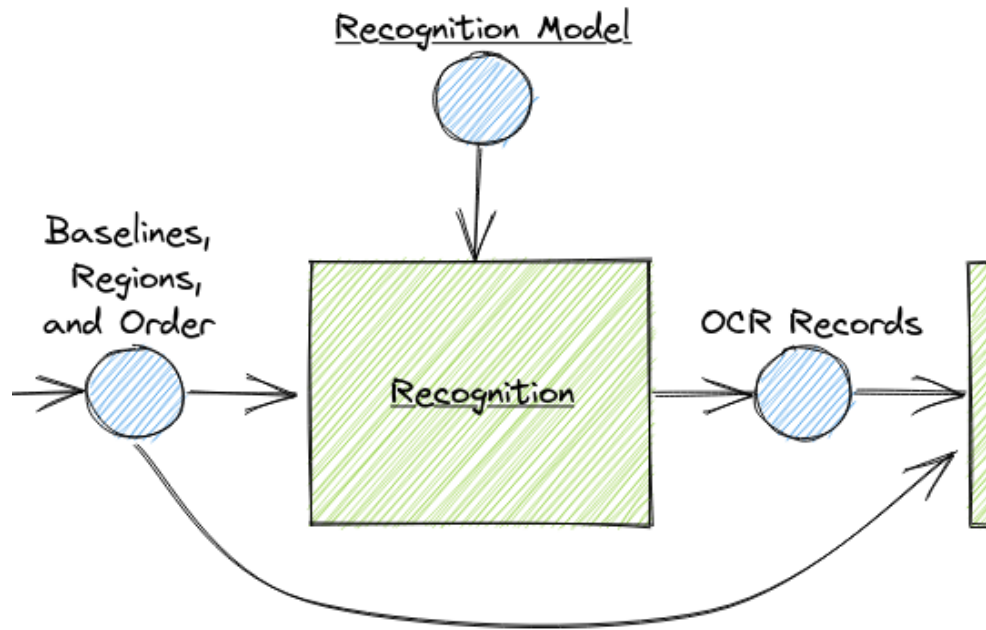
eScriptorium Home Kontakt AI Suche in 1800 VI bis 192 Meine Projekte Meine Modelle Hallo Thomas

Beschreibung Ontologie Bilder Bearbeiten Modelle Berichte 1800 VI bis 1920 IX manual

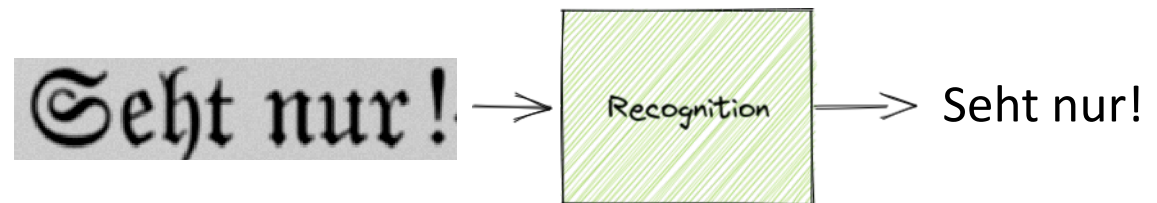
Zeilenmaske (line mask)
Grundlinie (baseline)
Textregion



Workflow-Schritt: Texterkennung

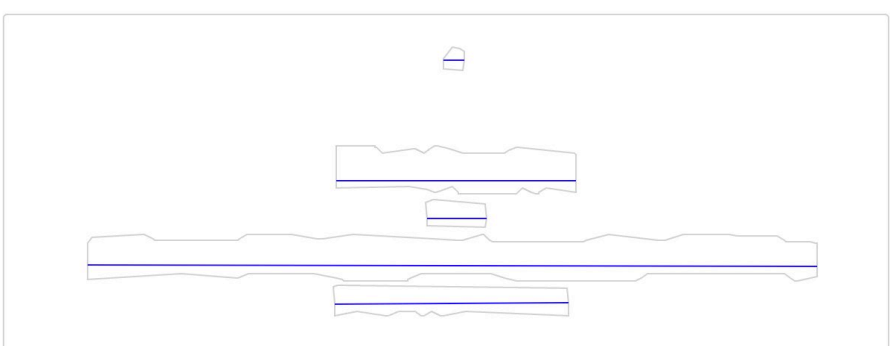


Bei den Texterkennungssysteme der letzten Jahre geht es darum, eine **Eingabe variabler Größe** (in der Regel auf der Grundlage eines Bilds einer Textzeile) zu haben und eine **Ausgabe derselben oder einer anderen Größe** (Gruppe von Glyphen) zu erzeugen.



Bildcredit: <https://kraken.re/main/index.html>



Workflow-Schritt: Texterkennung



17

Vorlagen
für die
Stadtverordneten-Verfammlung zu Charlottenburg.
In nicht öffentlicher Sitzung.

Vergleich eScriptorium und Transkribus

| | eScriptorium  | Transkribus  |
|--------------------|--|---|
| <i>Kosten</i> | Open Source, kostenfrei, skalierbar | Closed Source, Lizenzmodell mit steigenden Kosten bei vielen Nutzenden |
| <i>Datenschutz</i> | Daten verbleiben auf eigenen Servern | Server im Ausland oder On-Prem(ises) |
| <i>Support</i> | Aktive Entwicklerszene (Software-Updates, neue Funktionen) | Entwicklersupport (Ticketing-System mit Wartezeiten) |
| <i>Community</i> | Wachsende Community: Zahlreiche Anleitungen und Guides verfügbar | Große Community, weite Verbreitung, viele Anleitungen |
| <i>Funktionen</i> | Keine Tabellenstrukturen | Erfassung von Tabellenstrukturen |
| <i>FAIR Data</i> | Trainierte Modelle können heruntergeladen und geteilt werden | Trainierte Modelle können nicht heruntergeladen werden; Verwendung von Daten der Nutzenden! |

Mögliche Vorteile

- **Quelloffenheit:** jeder kann eine eigene Instanz installieren & zur Entwicklung beitragen
- Modelle und Ground Truth können Forschungsgemeinschaft zur Verfügung gestellt werden (**FAIR-Prinzipien**)
- **Wachsendes Angebot** von Basismodellen
- **Aktive Weiterentwicklung** mit State-of-Art Techniken
- Mittlerweile **breite** und **aktive Anwender*innenschaft**
- **Direkter Kontakt** mit den **Entwickler*innen** via Gitter
- bietet eine **moderne, webbasierte Benutzeroberfläche** zum **kollaborativen Arbeiten**
- **DSGVO:** Daten können im eigenen Kosmos bleiben, **kein Upload auf Fremdserver**
- **Kostenfaktor:** Kann langfristig günstiger sein als Transkribus

Mögliche Vorteile

- **Webapplikation:** keine Installation, kein Hosting, keine Updates nötig
- **All-in-One**, Ökosystem von Image bis zur Publikation
- Sehr **umfassende Modelle** können Trainingsaufwand reduzieren oder unnötig machen, auch bei **heterogenen Korpora**
- **Schnelle Weiterentwicklung**
- **On-Premise-Lösungen** erhältlich
- *Community-Unterstützung:* Workshops, Tagungen, Tutorials
- User-Management für **kollaboratives Arbeiten**
- **Annotationsmöglichkeiten**
- **API**

Vielen Dank für Ihre Aufmerksamkeit!

Larissa Will (larissa.will@uni-mannheim.de)

Jan Kamlah (jan.kamlah@uni-mannheim.de)

Stefan Büdenbender (stefan.buedenbender@h-da.de)

NFDI4Memory wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG);
Projektnummer 501609550.

BERD@NFDI wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG);
Projektnummer 460037581.