



Foto: Valentin Marquardt, Universität Tübingen

# Hands-On-Lab: Einführung in die Transkription von Handschriften und Drucken mit eScriptorium

BiblioCon 2023

Larissa Will, Stefan Weil (UB Mannheim)

Dorothee Huff (UB Tübingen)



## Vorhaben

---

Vorstellung des Kompetenzzentrums OCR der UBs Mannheim und Tübingen

---

Kurze Vorstellungsrunde – bisherige Erfahrungen mit OCR

---

Wie funktioniert automatische Texterkennung?

---

Workflow moderner Texterkennung

---

Was ist eScriptorium?

---

Schritt-für-Schritt-Einführung

---

Bearbeiten eigener Dokumente

---

Erfahrungsaustausch

---



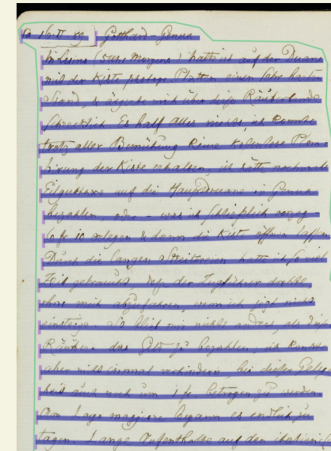
## Kompetenzzentrum OCR

- Entstanden aus dem Projekt OCR-BW
  - Laufzeit: 2019–2022
  - Projektpartner: UB Tübingen und UB Mannheim
  - gefördert durch: Ministerium für Wissenschaft, Forschung und Kunst
  - Aufgabe: Unterstützung von Archiven, Bibliotheken und anderen kulturbewahrenden Einrichtungen bei Anwendung von OCR- und Transkriptionssoftware
  - Ziel: Aufbau eines Kompetenzzentrums für automatisierte Texterkennung von Drucken und Handschriften



# Kompetenzzentrum OCR

- Kompetenzzentrum OCR mit Kooperationspartnern UB Tübingen und UB Mannheim bleibt weiterhin als Ansprechpartner bestehen
- Ausbau des Serviceangebots
- Fortführung bestehender Kooperationen, aber auch neue Projekte mit neuen Partnern
- Organisation eigener Veranstaltungen wie dem „Transcribathon durch den Orient“ in der Love Data Week 2023
- Offene Sprechstunde, jeden zweiten Donnerstag im Monat, online



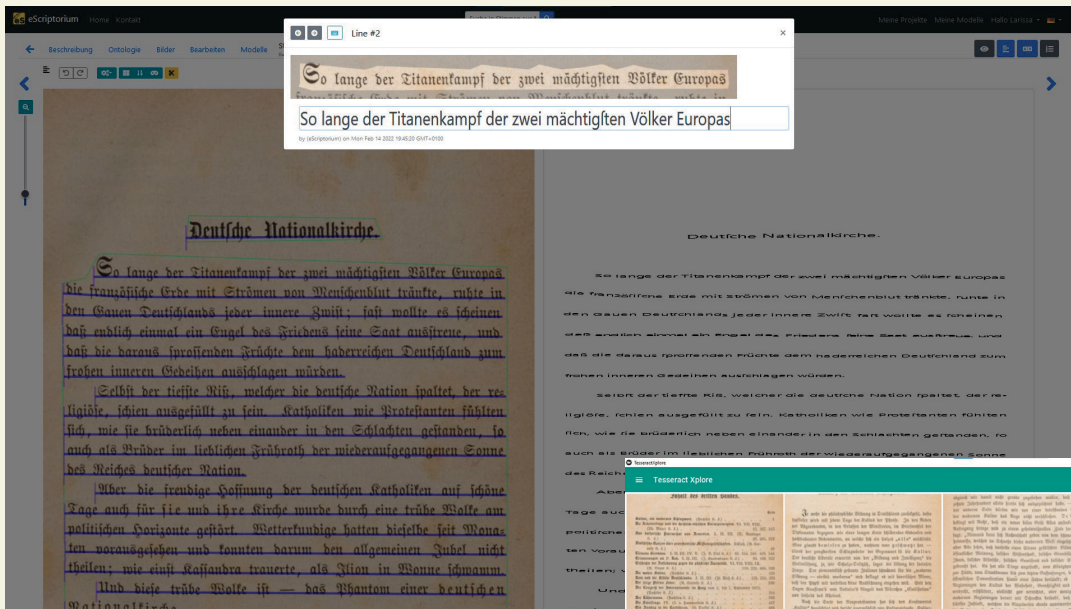
Sa 16.11.89 Gotthard-Genua

In Luins (5 Uhr Morgens) hatte ich auf der Duane mit der Kiste fotogr. Platten einen sehr harten Stand, & ärgerte mich über diese Räuberbande (schrecklich. Es half Alles nichts, ich konnte trotz aller Bemühung keine kostenlose Plombierung der Kiste erhalten. ich hätte nochmals Eilguttaxe auf die Hauptdouane in Genua bezahlen, oder - was ich schließlich vorzog - 6 fr. 10 erlegen & dann die Kiste öffnen lassen. Durch die langen Streitereien hatte ich so viel Zeit gebraucht, daß der Zugführer drohte, ohne mich abzufahren, wenn ich jetzt nicht einsteige. So blieb mir nichts anders, als diesen Räuber das Geld zu bezahlen; ich konnte aber nicht einmal verhindern bei dieser Gelegenheit auch noch um 1 fr. betrogen zu werden. Am Lago maggiore begann es endlich zu tagen. Lange Aufenthalte auf den italienisch Stationen bef. in Novara & Aleffandria. Die italien. Gendarmen mit ihren Dreimastern





# Bisherige Erfahrungen mit OCR?

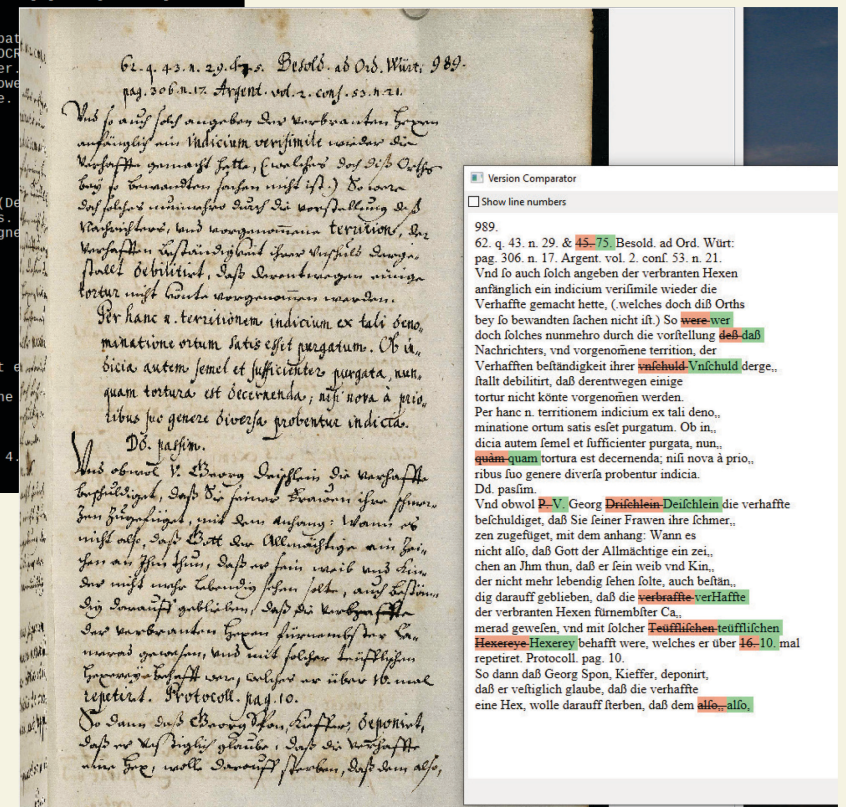
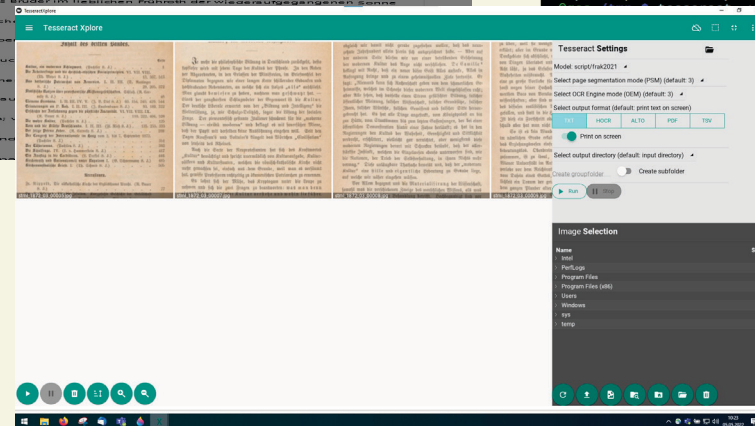


```
@geo /tmp $ tesseract
Usage:
tesseract imagename|stdin outputbase|stdout [options...] [configfile...]

OCR options:
--tessdata-dir /path specify location of tessdata path
-l lang[+lang] specify language(s) used for OCR
-c configvar=value set value for control parameter.
                    Multiple -c arguments are allowed
-psm pagesegmode specify page segmentation mode.
These options must occur before any configfile.

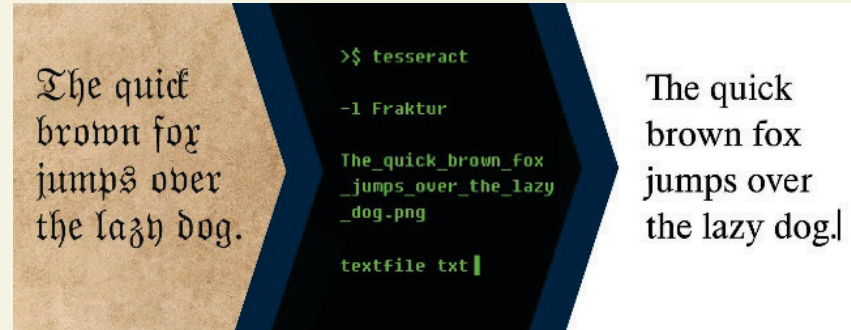
pagesegmode values are:
0 = Orientation and script detection (OSD) only.
1 = Automatic page segmentation with OSD.
2 = Automatic page segmentation, but no OSD, or OCR.
3 = Fully automatic page segmentation, but no OSD. (Default)
4 = Assume a single column of text of variable sizes.
5 = Assume a single uniform block of vertically aligned text.
6 = Assume a single uniform block of text.
7 = Treat the image as a single text line.
8 = Treat the image as a single word.
9 = Treat the image as a single word in a circle.
10 = Treat the image as a single character.

Single options:
-v --version: version info
--list-langs: list available languages for tesseract
--tessdata-dir:
--print-parameters: print tesseract parameters to the
```





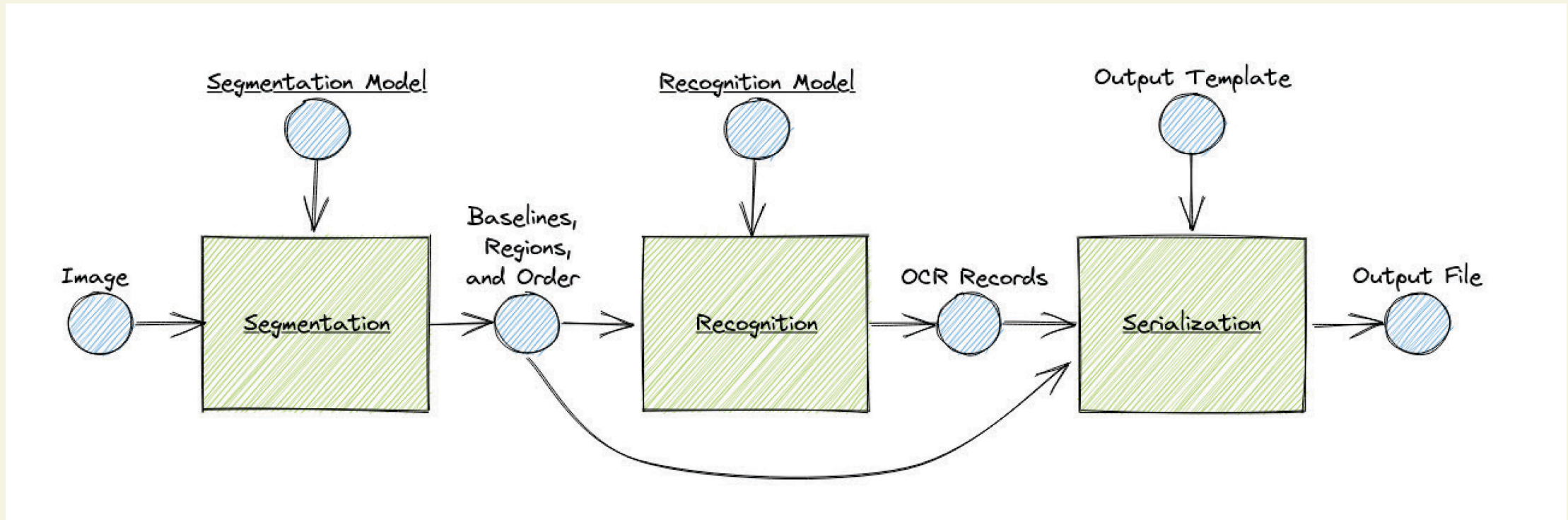
## Wie funktioniert automatische Texterkennung?



- Textliche Bildinhalte → digitale Textformate
- OCR: „Optical Character Recognition“ (optische Zeichenerkennung)
- Begriff mittlerweile veraltet
- Neuronale Netze erkennen nicht mehr Zeichen für Zeichen, sondern ganze Zeilen
- Texterkennung und OCR wird im deutschen Sprachraum oft synonym verwendet



# Workflow einer modernen Texterkennung





## Was ist eScriptorium?

- 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







## Schritt-für-Schritt-Einführung



Gehen Sie auf:

<https://ocr-bw.bib.uni-mannheim.de/escriptorium/>



Melden Sie sich mit den  
folgenden Zugangsdaten an:

Benutzername: bibliocon1

Passwort: Hannover23!



Wir gehen zusammen Schritt für Schritt durch eScriptorium!



## Weiterführende Ressourcen

### **Projekt OCR-BW/Kompetenzzentrum OCR**

<https://ocr-bw.bib.uni-mannheim.de>

### **eScriptorium**

<https://gitlab.com/scripta/escriptorium>

<https://ocr-bw.bib.uni-mannheim.de/escriptorium>

### **Modelle für eScriptorium/Kraken**

<https://ocr-bw.bib.uni-mannheim.de/faq/>

[https://zenodo.org/communities/ocr\\_models](https://zenodo.org/communities/ocr_models)

### **Mailingliste**

[https://listserv.uni-tuebingen.de/mailman/listinfo/ocr\\_htr\\_ub](https://listserv.uni-tuebingen.de/mailman/listinfo/ocr_htr_ub)

### **Kontakt**

Dorothee Huff

[dorothee.huff@uni-tuebingen.de](mailto:dorothee.huff@uni-tuebingen.de)

Larissa Will

[larissa.will@uni-mannheim.de](mailto:larissa.will@uni-mannheim.de)

Stefan Weil

[stefan.weil@uni-mannheim.de](mailto:stefan.weil@uni-mannheim.de)