# Volltexterkennung für historische Sammlungen mit OCR4all-libraries iterativ und partizipativ gestalten

## Nutzen |

- OCR für den eigenen Kontext je nach Sammlungsanforderungen und Forschungsfragen flexibel und eigenständig einsetzbar
- Verbesserung der OCR-Prozesse und Qualität für Sammlungen mit unterschiedlichen Layouts und Drucktypen
- OCR als iterativer und parzipativer Prozess, bei dem Wissenschaftler:innen aus den Digital Humanities beteiligt und als Nutzer:innen eingebunden werden


gei-digital.gei.de

Jan Sebastian Klaes (GEI), Kristof Korwisi (ZPD), Katharina Krüger (GEI), Christian Reul (ZPD), Nadine Towara (GEI)

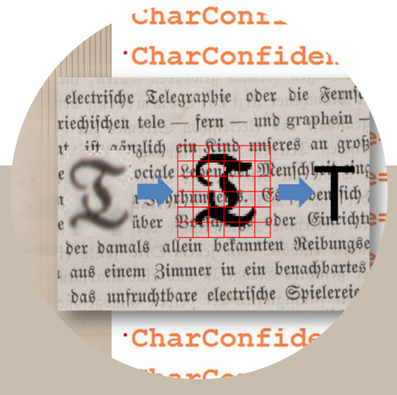
Im Zentrum der Zusammenarbeit steht die Erweiterung und Anpassung des GUI-basierten Open-Source-Werkzeugs OCR4all, sodass Bibliotheken und Archive bei ihrer Massendigitalisierung die im Rahmen des OCR-D-Projekts erarbeiteten Lösungen niederschwellig, flexibel und eigenständig einsetzen können.

## Use Case |

- Digitalisierte historische Lehrbuchsammlungen des GEI mit der stark variierender OCR-Qualität

## OCR4all |

- Open Source
- Nutzungsfreundlich GUI mit intuitiver Benutzung
- Eigene Modelle für die Texterkennung einfach trainieren und anwenden



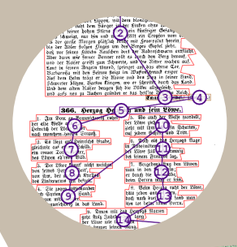

## OCR |

- OCR4all stellt einen umfangreichen und hochkonfigurierbaren OCR Workflow zur Verfügung

Vorverarbeitung

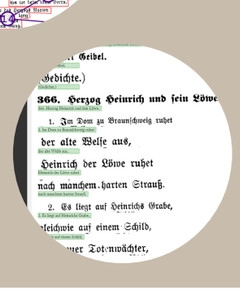
Layoutanalyse

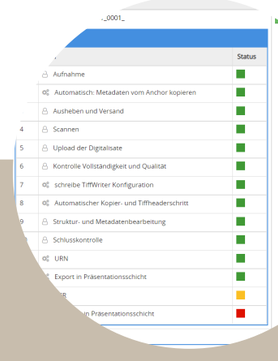
Erkennung



Korrektur

Training




## Ziele |

- Out-of-the-box-Ansatz für Anbindung von OCR4all in den Goobi-Digitalisierungsworkflow
- Studie zur Integration von OCR4all in Digitalisierungssysteme
- Images und Volltexte direkt aus dem Goobi-Viewer downloaden und in OCR4all integrieren

## GEI|DIGITAL

- 5.600 historische deutsche Schulbücher von den Anfängen der Schulbuchproduktion bis 1918
- Lehrbücher aus den Fächern Geschichte, Geographie und Politik sowie Realien- und (Erst)-Lesebücher
- Über 1,5 Millionen Digitalisate
- Differenzierte Inhaltserschließung mit 143.000 Strukturelementen und 326.000 Metadaten
- Umfangreiche Volltexterkennung mit OCR (ALTO/ABBY XML)
- Metadaten und Volltexte über OAI-PMH-Schnittstelle abrufbar

LEIBNIZ-INSTITUT FÜR BILDUNGSMEDIEN |Georg-Eckert-Institut

Zentrum für Philologie und Digitalität Kallimachos

Leibniz Gemeinschaft