Zweitveröffentlichung

Scheltjens, Werner; Geschonke, Sebastian; Gerstner, Jan; u. a.

Aufbau und Bereitstellung eines Benchmark Datensatzes für Historische Tabellen, 1750-1990

Datum der Zweitveröffentlichung: 28.09.2026

Verlagsversion (Version of Record), Konferenzveröffentlichung

Persistenter Identifikator: urn:nbn:de:bvb:473-irb-117343x

Erstveröffentlichung

Scheltjens, Werner; Geschonke, Sebastian; Gerstner, Jan; u. a. (2026): Aufbau und Bereitstellung eines Benchmark Datensatzes für Historische Tabellen, 1750-1990, Zenodo, doi: 10.5281/zenodo.22861458.

Rechtehinweis

Dieses Werk ist durch das Urheberrecht und/oder die Angabe einer Lizenz geschützt. Es steht Ihnen frei, dieses Werk auf jede Art und Weise zu nutzen, die durch die für Sie geltende Gesetzgebung zum Urheberrecht und/oder durch die Lizenz erlaubt ist. Für andere Verwendungszwecke müssen Sie die Erlaubnis der Rechteinhaberinnen und Rechteinhaber einholen.

Für dieses Dokument gilt eine Creative-Commons-Lizenz.

Die Lizenzinformationen sind online verfügbar:

https://creativecommons.org/licenses/by/4.0/legalcode

NFDI Community Forum
Bonn, 17. September 2025

Universität Bamberg

Aufbau und Bereitstellung eines Benchmark Datensatzes für historische Tabellen, 1750-1990

Werner Scheltjens, Sebastian Geschonke, Jan Gerstner, Isabel Rösner, Konstantina Slavova

I. Relevanz und Ziele

Trotz rasanter Fortschritte in der KI-gestützten Texterkennung historischer Drucke bleibt die **Erkennung historischer Tabellen** eine komplexe Herausforderung. Bislang fehlen allgemeine Lösungsvorschläge, in denen die Eigenschaften und Eigentümlichkeiten historischer Tabellen bei der KI-gestützten Struktur- und Inhaltserkennung berücksichtigt werden.

Tabellenerkennung in 3 Fragen
1. Wo befindet sich die Tabelle?
2. Welche Struktur hat die Tabelle?
3. Welche Inhalte hat die Tabelle?

Ziele: 1. Aufbau eines repräsentativen Datensatzes von ca. 10,000 ganz- und mehrseitigen gedruckten historischen Tabellen im Bereich der **Sozial- und Wirtschaftsgeschichte** ab ca. 1750; 2. Bereitstellung des Datensatzes mit Labels und Annotationen, als Grundlage für die Entwicklung von neuen Lösungen für die Herausforderung der automatisierten historischen Tabellenerkennung.

II. Herausforderungen und Lösungen

1. Tabellen finden

Problem	Tabellenseiten sind in digitalen Sammlungen nicht ohne weiteres auffindbar, weil sie nicht in den Metadaten verzeichnet werden.
Lösung	1. Erfassung der Metadaten von Digitalisaten mit den Signaturfächern **Merc., Oekon., Cam. & Num.** und Varianten der Wörter **‚Tabelle‘** oder **‚Statistik‘** im Titel über die **OAI-PMH** Schnittstelle der BSB München (ergänzend StaBi Berlin). 2. Bislang erfasster Zeitraum: **1750-1900**. 3. Speicherung der Metadaten in einer **MongoDB**. 4. Training eines Klassifikationsmodells mit **Keras** für Scans historischer Drucke (Ziel: Erkennung von Tabellenseiten). 5. Aufruf der Bilder mit **IIIF** aller in der Datenbank erfassten Digitalisate. ML-basierte Klassifikation jedes Bildes. 6. Anwendung des Klassifikationsmodells. 7. ‚Tabellenseiten‘ werden als JPG und PNG gespeichert.

2. Datensatz zusammenstellen

Problem	Mehr als 100,000 Scans aus >1,500 Büchern. Manchmal nur 1 Bild; manchmal > 1,000 Bilder pro Buch.
Lösung	1. Kriterienbasierte Auswahl: Confidence Rate 1.00; Max. 200 Seiten pro Buch; Min. 5 und Max 100 Tabellenseiten pro Buch. 2. Weiterbearbeitung von **11,995 Scans aus 364 Büchern**.

3. Tabellen segmentieren

Problem	Historische Tabellen haben häufig eine komplexe, unregelmäßige und verschachtelte Struktur.
Lösung	1. Grobsegmentierung in **Label-Studio**: Labelling von **Attributes** (~Spaltenüberschriften), **Instances** (~Zeilenüberschriften) und **Values** als Bereiche auf (fast) jeder Tabellenseite. 2. Feinsegmentierung und Bounding Box Recognition mit einem eigenen Tool (**table preditor**), das die Layouterkennung für jedes identifizierte Grobsegment durchführt und manuelle Nachbearbeitung zum Zweck der Ground Truth Produktion sowie Annotation einzelner Bounding Boxes ermöglicht.

4. Tabellen annotieren

Problem	Die komplexe Struktur lässt sich nicht immer genau in den drei Großbereichen erfassen.
Lösung	Beschreibung historischer Tabellenseiten auf der Grundlage eines ausführlichen **Fragebogens** in **Label-Studio**. Antworten werden als JSON gespeichert.

5. Ground Truth produzieren

Problem	Weil es noch keine zuverlässige automatisierte Tabellenerkennung für historische Drucke gibt, fehlt auch Ground Truth.
Lösung	**OCR** der nachjustierten Feinsegmente mit **Anthropic API**. Manuelle Korrektur.

III. Nächste Schritte

- Finalisierung und **Veröffentlichung des Datensatzes** mit allen Koordinaten der Grob- und Feinsegmente und den korrigierten OCR-Ergebnissen.
- Bereitstellung von **Code und Data** auf **https://github.com/dhofu**

mittelfristig	Datensatz erweitern um Tabellen aus anderen Wissensbereichen
	Algorithmen für die Einzelschritte optimieren und weiterentwickeln
langfristig	Training einer KI-gestützten Lösung anstreben

Kontakt: Prof. Dr. Werner Scheltjens, Professur für Digitale Geschichtswissenschaft, Otto-Friedrich-Universität Bamberg, werner.scheltjens@uni-bamberg.de 0000-0002-5209-9052