ZfdG

Zeitschrift für digitale Geisteswissenschaften

Fachartikel aus:
Zeitschrift für digitale Geisteswissenschaften, Heft 10 (2025)

Titel:
Semiautomatische Erschließung von Fotografien auf beschrifteten Bildkarten im Archiv. Dokumentenerkennung mit Deep Learning sowie Large-Language-Modellen

Autor*in:
Norbert Fischer

Kontakt: norbert.fischer@informatik.uni-wuerzburg.de
Institution: Julius-Maximilians-Universität, Würzburg
GND: 130344230 ORCID: 0000-0002-7365-8239

Autor*in:
Dominik Kimmel

Kontakt: dominik.kimmel@leiza.de
Institution: Leibniz Zentrum für Archäologie
GND: 1047560313 ORCID: 0000-0001-5284-668X

Autor*in:
Frank Puppe

Kontakt: frank.puppe@uni-wuerzburg.de
Institution: Julius-Maximilians-Universität, Würzburg
GND: 17410085X ORCID: 0000-0002-7106-3223

DOI des Beitrags:
10.17175/2025_009

Nachweis im OPAC der Herzog August Bibliothek:
193053003X

Erstveröffentlichung:
28.08.2025

Letzte Überprüfung aller Verweise:
22.08.2025

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Archiv | Bilderkennung | Archäologie | Bildkarte | Optische Zeichenerkennung | Großes Sprachmodell

Empfohlene Zitierweise:
Norbert Fischer / Dominik Kimmel / Frank Puppe: Semiautomatische Erschließung von Fotografien auf beschrifteten Bildkarten im Archiv. Dokumentenerkennung mit Deep Learning sowie Large-Language-Modellen. In: Zeitschrift für digitale Geisteswissenschaften 10 (2025). 28.08.2025. HTML / XML / PDF. DOI: 10.17175/2025_009

Norbert Fischer / Dominik Kimmel / Frank Puppe

Semiautomatische Erschließung von Fotografien auf beschrifteten Bildkarten im Archiv. Dokumentenerkennung mit Deep Learning sowie Large-Language-Modellen

Abstract

Die automatisierte Extraktion von Bild-Textarchivalien kann deren Erschließung und FAIRe Bereitstellung effizienter machen und dadurch ihre Auffindbarkeit und Nachnutzung wie semantische Suchen oder datenbasierte Auswertungen verbessern. Wir berichten über ein Experiment, bei dem die Texte von beschrifteten Bildkarten des Leibniz-Zentrums für Archäologie (LEIZA) mit zwei Methoden erschlossen wurden: mit einer klassischen *Deep-Learning*-Pipeline sowie mit *Large-Language-Modellen (LLMs)* wie GPT-4o. Beide Ansätze erreichten gute Ergebnisse, wobei die *LLMs* besser mit den unterschiedlichen Strukturen der Bildkarten umgehen können. Mit automatisierter Vor– und Nachbearbeitung erzielten wir Erkennungsraten von über 90 %. Herausforderungen sind der Umgang mit unterschiedlichen Bildkarten-Strukturen und handschriftlichen Einträgen, bei denen die Extraktion deutlich schlechter funktioniert.

The automated extraction of image-text archival material can make their cataloguing and FAIR provision more efficient, improve their findability and subsequent uses such as semantic search or data-based analyses. We report on an experiment in which the texts of record cards from the Leibniz-Zentrum für Archäologie (LEIZA) were semantically extracted using two methods: a classical *deep learning* pipeline and *large language models (LLMs)* such as GPT-4o. Both approaches achieved good results, with the *LLMs* being better able to deal with the different structures of the record cards. With automated pre– and post-processing, we achieved recognition rates of over 90 %. Challenges comprise the different record card structures and handwritten entries, for which the extraction performs significantly worse.

1. Ausgangslage und Fragestellung

In vielen Museen und Archiven befinden sich große Mengen an Fotografien, die noch nicht -oder nur flach, mit wenig Informationstiefe -digital erschlossen sind. Analoge fotografische und grafische Archivalien werden in der Archivpraxis in der Regel noch immer in aufwändiger händischer Arbeit erschlossen. Automatisierte, algorithmusgestützte Analysen von Bildquellen stehen zwar schon länger im Fokus der Digital Humanities, zur automatisierten Erschließung und Analyse größerer Mengen von Fotografien bestehen aber erst seit jüngster Zeit praxistaugliche Ansätze.[1] Eine Reihe von vielversprechenden Herangehensweisen befindet sich im Experimentstatus oder wird derzeit für die Praxis historischer oder Kulturerbe-bezogener Archivalien erprobt.[2] [1]

Für eine automatisierte Erschließung besonders vielversprechend erscheinen derzeit Bildarchivalien, zu denen auch Textinformationen vorliegen. Die Texte in Beschriftungen auf oder bei den Fotos selbst oder aber in begleitenden Materialien enthalten oftmals wesentliche Informationen zu den Archivalien und den dort abgebildeten Motiven, die nicht nur der Beschreibung des Sichtbaren oder der Klassifikation dienen, sondern vielfach auch Daten zur Provenienz und andere inhaltliche Erschließungsinformationen enthalten. Eine automatisierte multimodale Extraktion und strukturierte Aufbereitung der Texte dieser Archivalien kann deren Erschließung und FAIRe Bereitstellung deutlich effizienter machen und ihre Auffindbarkeit sowie Nachnutzungen wie semantische Suchen oder datenbasierte Auswertungen auch sammlungsübergreifend verbessern. [2]

[1] Vgl. u. a. Wevers / Smits 2020; Smits / Wevers 2023; Arnold / Tilton 2023.

[2] Die Tagung ›Artificial Intelligence in Archives and Collections: Practices, Potentials, and Evidence Production in Dealing with Images and Multimodal Cultural Heritage‹ vom 12.–13.12. des Leibniz-Forschungsverbundes Wert der Vergangenheit hat hier wesentliche Aspekte der aktuellen Diskussion zusammengeführt: Tagungswebsite. Die Zusammenfassung wird 2025 bei Hypotheses veröffentlicht: Value of the Past.

[3] Der vorliegende Bericht stellt Ergebnisse eines Experiments vor, das die Autoren im Rahmen ihrer Untersuchungen zu den Einsatzmöglichkeiten geeigneter Technologien für die Archiv– und Sammlungspraxis durchgeführt haben. Er befasst sich insbesondere mit den Herausforderungen, die Texte von Bildarchivalien zu erschließen. In diesem Beitrag konzentrieren wir uns auf deren Extraktion in strukturierter Form, sodass die Informationen im Anschluss beispielweise direkt in Datenbanken oder Wissensgraphen überführt werden können. Wir haben dazu zwei verschiedene Ansätze erprobt: klassische *Deep-Learning*-Ansätze zur Dokumentenanalyse wie Layout-Analyse und *Optical Character Recognition (OCR)* und *Handwritten Text Recognition (HTR)*, sowie neue *multimodale Large Language Modelle (LLMs)* wie GPT-4o.

[4] Als Use Case dient die Sammlung an Bildkarten mit beschreibendem Text des Bildarchivs des Leibniz-Zentrums für Archäologie (LEIZA), das 1852 als Römisch-Germanisches Zentralmuseum (RGZM) gegründet wurde und bis 2022 diesen Namen trug. Das heutige LEIZA-Archiv besitzt etwa 150.000 Bildkarten mit rund 200.000 Abbildungen aus der Zeit von etwa 1890 bis 2005.

2. Datengrundlage: Bildkarten des Leibniz-Zentrums für Archäologie

[5] Zu den Bild-Text-Archivalien, die im LEIZA, wie auch in anderen Sammlungen in großer Menge vorliegen, gehören sogenannte Bildkarten: Kartonblätter, auf denen ein oder mehrere Abbildungen – ein Foto, eine Grafik oder ein Druck – gemeinsam mit Verzeichnungsinformationen oder Beschriftungen aufgezogen wurden. Diese Karten wurden seit dem frühen 20. Jahrhundert angelegt und in einer archiveigenen, fachspezifischen Systematik geordnet, beispielsweise nach Orten, Zeiten oder inhaltlichen Kriterien, und so aufgestellt, dass sie für wissenschaftliche Zwecke oder andere Recherchen zugänglich und zu finden waren.

[6] Da sie über einen Zeitraum von rund 100 Jahren angelegt wurden, haben die Bildkarten in der Fotosammlung des LEIZA keine einheitliche Gestalt und Anordnung von Bildern und Texten. Die Textinformationen liegen getippt, gestempelt und handschriftlich in unterschiedlichsten Layouts und Schrifttypen vor. Die Karten dienten aber immer dazu, Abbildungen von archäologischen Objekten, Befunden und auch anderen Motiven aus der Arbeit des Instituts systematisch abzulegen und für die wissenschaftliche Arbeit institutsinterner aber auch externer Forschender auffindbar zu machen. Aus diesem Grund wurden über lange Zeiträume gleichartige Kartons verwendet und die Karten sind zumeist, soweit das über die lange Zeit möglich war, von gleicher oder ähnlicher Größe. Sie tragen, selbst wenn sich Layout, Schrift und Gestalt über die Zeit geändert haben, meist strukturell dieselben oder gleichartige Informationen, die sich über die Zeit nur wenig verändert haben. Meist finden sich folgende Angaben zu den auf die Bildkarte aufgezogenen Abbildungen (›Inhaltsschema‹):

- Ortsbezug des Motivs (meist Fundort eines archäologischen Objektes oder Befundes): in unterschiedlicher Detailliertheit zu Fundstelle und geografischer Verortung des Fundortes.
- Gegenstand: Ansprache, kurze inhaltliche Beschreibung des Motivs, in den meisten Fällen ein Objekt oder auch ein Architekturbestandteil oder archäologischer Befund.
- Zeitliche Einordnung des Motivs: grobe Zeitangabe, z. B. in Form eines Jahrhunderts oder einer Periode wie frühawarisch oder römisch.
- Verwahrer des Motivs (typischerweise eines Objektes): meist ein Museum.
- Literatur: Angabe einer Literaturstelle unter der das abgebildete Objekt gezeigt oder erwähnt wird; in manchen Fällen die Vorlage eines Reprofotos.
- Registrierungsnummern: eine oder mehrere Buchstaben-Zahlen-Kombinationen, meist Inventarnummern der Objekte oder Bildnummern.

Innerhalb dieser grundsätzlichen inhaltlichen Struktur lassen sich für den Hauptbestand der Karten verschiedene Grade der Strukturiertheit feststellen, die für die automatisierte Transkription einer gesonderten Betrachtung bedürfen. [7]

Eine erste Gruppe von Bildkarten wurde nach einer vorgegebenen Systematik beschriftet, d. h. einem Schema mit vorgegebenen Kategorien, das bei der Archivierung des Bildes ausgefüllt oder ausgedruckt und dann auf den Trägerkarton geklebt wurde. Diese ›schematischen‹ Bildkarten haben den höchsten Grad der Strukturierung und sind daher am besten transkribierbar. Das zugrundeliegende Formular umfasst folgende zehn Kategorien: [8]

- Zum Ortsbezug des Motivs: (1) FO (Fundort), (2) Fdst (Fundstelle), (3) Kreis, (4) Land.
- Zum Verwahrer des Motivs: (5) Mus (Museum). Der Eintrag verweist auf das Bild eines Objektes, das sich nicht im LEIZA befindet. Teilweise werden hier auch die Objekt-Inventarnummern dieser Verwahrer, meist Museen, angeführt. Hier kann es sich auch um den Verwahrort des Originals zu einem Objekt handeln, das sich in Kopie in der LEIZA-Objektsammlung befindet.
- Zu eindeutigen Registrierungsnummern: (6) RGZM, (7) Neg (Negativ). Unter (6) ›RGZM‹ wird die Inventarnummer eines Objektes angeführt, wenn es sich in der Sammlung des LEIZA befindet. Unter (7) ›Negativ‹ wird die eindeutige Negativnummer des gegenständlichen Fotos angeführt. Dieses kann entweder von LEIZA-eigenen Fotografen oder von externen Fotografen angefertigt worden sein. Die LEIZA-internen Fotos werden am Zusatz ›RGZM‹ oder an bestimmten Präfix-Nummern-Kombinationen (z. B. T 2001/853) erkannt.
- Zum Gegenstand / Beschreibung des Motivs: (8) Ggst.
- Zur zeitlichen Einordnung des Motivs: (9) Zeit
- Zur Literatur: (10) Lit.

Abb. 1: Verschiedene Schemata der Bildkarten im LEIZA-Archiv. Oben: Zwei schematisch beschriebene Bildkarten basierend auf einem – teilweise vorgedruckten – ›Formular‹, dessen Einträge mit Schreibmaschine ausgefüllt wurden (links: einfaches Layout, rechts: kompliziertes Layout); unten: zwei teil-schematisch beschriebene Bildkarten, bei denen nur manche Kategorie-Namen des obigen ›Formulars‹ angegeben sind. [Fotos: unten links: Simone Deyts / Claude Rolley: L'Art de la Bourgogne romaine, découvertes récentes. Musée archéologique de Dijon, France 1973, Kat. Nr. 25; unten rechts: Mainzer Zeitschrift 36, 1941 Taf. II,1; alle LEIZA-Archiv]

Die inhaltlichen Angaben des Formulars sind in der Regel mit Schreibmaschine geschrieben und teilweise auch handschriftlich ergänzt oder korrigiert. Oft fehlen Angaben zu einigen Kategorien. Zwei Beispiele zeigt Abb. 1. Eine Herausforderung besteht – neben den handschriftlichen Einträgen – darin, dass die inhaltlichen Angaben oft nicht in den vorgesehenen Platz des Formulars gepasst haben und dann in vorhandenen freien Raum weitergeschrieben wurden, d. h. in den Platz einer anderen Kategorie (Abb. 1, oben rechts). [9]

Eine weitere Gruppe von Bildkarten nutzt kein strukturiertes Formular, verwendet aber teilweise dessen Begriffe und eine grundsätzlich ähnliche Strukturierung (Abb. 1, unten). Hier wird die Zuordnung der inhaltlichen Texte zu dem Schema erschwert, da entweder keine Schemawörter angegeben sind (z. B. beim Gegenstand) oder nur teilweise (z. B. beim Fundort). Außerdem ist die Variationsbreite der Struktur der Texte deutlich heterogener. Diese zweite Gruppe von Bildkarten nennen wir ›teil-schematisch‹. [10]

Eine dritte Gruppe von Bildkarten enthält keine Schemabegriffe und wird daher als ›nicht-schematisch‹ klassifiziert. Aber auch in dieser Gruppe sind grundsätzlich die gleichen Informationen auf der Bildkarte enthalten. Die Transkription wird zusätzlich dadurch erschwert, dass bei manchen Karten wichtige [11]

Informationen als Text im Bild selbst enthalten sind und nicht in dem beschreibenden Text unter der Abbildung. Daher ist diese Gruppe der nicht-schematischen Bildkarten deutlich heterogener als die ersten beiden Gruppen.

[12] Darüber hinaus gibt es eine Reihe von Bildkarten, die in keines der oben genannten Schemata passen, da sie zumeist Fotos zeigen, die nicht von archäologischen Objekten, Architektur oder Befunden stammen. Sie zeigen beispielsweise Personen, Fotos von Ausstellungen oder andere Bilder mit Bezug zur Arbeit des ehemaligen Römisch-Germanischen Zentralmuseums.

[13] Zusätzlich tragen viele Bildkarten am rechten oder oberen Rand eine Signatur, die einen Teil der Schemainformationen kodiert (Abb. 2, links). Sie besteht meist aus drei Ziffernblöcken mit oder ohne Schrägstriche zur Trennung, teilweise mit textueller Zusatzinformation. Die Signaturen kodieren wesentliche Informationen zum Motiv (grobe zeitliche Einordnung, Fundort, für den römerzeitlichen Teil der Bildkarten auch eine grobe semantische Klassifizierung der Motive), die bis auf die semantische Klassifizierung der römischen Motive auch in der inhaltlichen Beschreibung auf den Karten enthalten sind. Die Signatur-Informationen sind also teilweise redundant. Sie sind in sich relevant, können aber auch zur automatischen Qualitätskontrolle genutzt werden. Auch zur Erkennung der Signaturen sowie der schematischen und teil-schematischen Bildkarten wurden Experimente durchgeführt, deren Ergebnisse in Kap. 4 präsentiert werden.

Abb. 2: Signaturen auf Bildkarten des LEIZA-Bildarchivs. Beispiele für drei gestempelte Signaturen (links), eine handschriftliche Werkblatt-Nummer unter dem nur rudimentär ausgefüllten Signaturfeld (Mitte) und eine handschriftliche Signatur in einem gestempelten blauen Kästchen (rechts). [Ausschnitte aus Bildkarten, LEIZA-Archiv].

[14] Es gibt aufgrund der gelebten Praxis und des langen Zeitraums der Erfassung viele Besonderheiten wie u. a. nachträgliche handschriftliche Korrekturen und Ergänzungen, Stempel, die Zusammenfassung mehrerer Bildobjekte in einer Bildkarte oder das Hinzufügen von Registriernummern am Rand. Ein Teil der Karten enthält an nicht genau definierter Stelle sogenannte Werkblattnummern, die auf die Dokumentation konservatorisch-restauratorischer Arbeiten, die im LEIZA an den abgebildeten Objekten durchgeführt wurden, verweisen (Abb. 2, Mitte). Weiterhin finden sich auch Signaturen, die in einem blauen Kästchen handschriftlich eingetragen sind und gleichartige Informationen wie die Signaturen links kodieren, aber in einem anderen Format (Abb. 2, links). Darüber hinaus tragen die Bildkarten teilweise einen roten Stempel »Veröffentlichung nur mit Genehmigung des Eigentümers der Gegenstände gestattet« (Abb. 1, oben rechts), der sich meist am linken Rand befindet, aber teilweise über dem strukturellen Formular gestempelt wurde, wodurch die Transkription der hier vermerkten Informationen zusätzlich erschwert wird. Eine weitere Herausforderung für die Transkription ist die unterschiedliche Anzahl an Bildern, die auf einem Karton präsentiert und beschrieben wurden. Oft handelt es sich um mehrere Abbildungen zu einem archäologischen Objekt, teilweise werden aber auch zusammenhängende Funde aus mehreren Objekten anhand verschiedener Fotos auf einer Karte abgebildet.

3. Methoden

3.1 Herausforderungen und grundsätzliche Vorgehensweise

[15] Zur multimodalen Erschließung der Bildkarten (vgl. Abb. 1) sind folgende Herausforderungen zu lösen:

- Erkennung des Kartentyps / Layouttyp der Karte

- Segmentierung: Trennung von Text– und Bildregionen
- Erkennung des Layouts der Textregionen und der Bedeutung der Felder
- Transkription der gedruckten und handschriftlichen Texte mit *OCR*– und *HTR*-Modellen
- Zuordnung der extrahierten Texte zu sinnvollen Datenbankfeldern
- Optional: Durchführung automatischer Korrekturen mit Hintergrundwissen

Dabei sollen aus den Texten der Bildkarten die oben genannten zehn Kategorien extrahiert werden, die im Allgemeinen nur teilweise angegeben und bei den schematischen Bildkarten explizit, bei den teil-schematischen teils explizit und teils implizit und bei den nicht-schematischen nur implizit enthalten sind. [16]

Für die Extraktion der strukturierten Informationen wurden zwei Lösungsansätze getestet. Der erste Ansatz basiert auf einer *OCR*-Pipeline bestehend aus Vorverarbeitung, Textdetektion, Texterkennung und anschließender Einordnung der erkannten Texte gemäß dem zu erkennenden Schema. Der zweite Ansatz benutzt GPT-4o als Ende-zu-Ende-Lösung für die Aufgabe. [17]

Für beide Lösungsansätze wurde eine automatisierte Vorverarbeitungs– und Nachbearbeitungskomponente hinzugefügt, wobei letztere Hintergrundwissen nutzt: Zum einen wird während der Transkription eine Datenbank mit Inhalten aus der Extraktion aufgebaut, da die Namen vieler Fundorte, Museen, Gegenstands– und Zeitangaben in verschiedenen Bildkarten vorkommen und auch die Registrierungsnummern eine reguläre Struktur haben. Zum anderen adressiert sie typische Fehler, wie beispielsweise längere Texte, die in frei verfügbaren Platz geschrieben werden, was syntaktisch (vom Layout) zu Fehlzuordnungen führt, aber meistens semantisch (durch den Inhalt der Texte) korrigiert werden kann. [18]

3.2 Vorverarbeitung der Eingabedaten

Die Ausgangsdaten für die Digitalisierung sind Scans der Bildkarten, welche als Bilddateien vorliegen. Die Scans können stellenweise noch schwarze Ränder enthalten, wenn diese nicht automatisch während des Scannens entfernt wurden. Da zudem die Bildkarten aus unserem Datensatz sowohl im Hoch– als auch Querformat vorliegen können (je nach Größe und Seitenverhältnis der enthaltenen Fotografien) und bei der Digitalisierung nicht strikt auf die korrekte Ausrichtung geachtet wurde, ist der erste Schritt, die Bildkarten korrekt zuzuschneiden und die Orientierung zu vereinheitlichen. Die Erkennung der Ränder wurde durch einen Algorithmus gelöst, welcher dunkle Bereiche am Rand abtrennt, aber robust gegenüber Rauschen ist. Für die Korrektur der Orientierung wurde die unten beschriebene *OCR*-Pipeline verwendet. Da die verwendeten Modelle ausschließlich bei korrekt orientierten Textzeilen funktionieren, wurde die Anzahl an gefundenen Kategorie-Wörtern bzw. die Länge der erkannten Texte (falls keine Kategorie-Wörter gefunden wurden) als Heuristik für die korrekte Orientierung herangezogen. [19]

3.3 Klassische Erkennungs-Pipeline

Die klassische Erkennungs-Pipeline basiert u. a. auf Vorarbeiten der Verfasser[3] und wurde anschließend noch optimiert. Im ersten Schritt werden hierbei einzelne Textzeilen mittels DBNet erkannt, wobei wir für unsere Experimente die vortrainierten DBNet-Modelle von Haofu Liao et al. verwenden.[4] Aus der Ausgabe von DBNet werden mithilfe der Implementierung aus DocTR[5] für einzelne Wörter achsenorientierte Rechtecke extrahiert. Nahe nebeneinander liegende Bounding-Boxen einzelner erkannter Wörter, welche zudem auf ähnlicher Höhe liegen, werden dabei zu einzelnen Textzeilen kombiniert, wobei darauf geachtet wurde, dass die beiden [20]

[3] Vgl. Fischer / Hartelt / Puppe 2023; Kammleiter 2024.
[4] Vgl. Liao 2023.
[5] Vgl. Liao 2023.

Spalten (s. Abb. 1, oben) nicht verbunden werden. Die zusammengesetzten Zeilen werden im Anschluss mittels eines *Deep Recurrent Neural Networks* für die *OCR* transkribiert. Die verwendete Netzwerkarchitektur ist dabei stark angelehnt an das *CRNN*[6] mit einzelnen Parameteranpassungen.

[21] Im letzten Schritt werden aus den erkannten Textzeilen die Kategorien mit ihren Einträgen extrahiert. Um dies zu erreichen, werden zuerst Textzeilen identifiziert, die die vordefinierten Kategorien des Schemas enthalten. Die übrigen Textzeilen werden anschließend der horizontal bzw. vertikal nächsten Kategoriezeile hinzugefügt. Schließlich werden die erkannten Einträge auf Basis einfacher Regeln nachgebessert, um bei ungenauer Positionierung der Einträge die Erkennungsgenauigkeit zu verbessern.[7] Dieses Verfahren benötigt Kategorie-Wörter und eignet sich daher nur für schematische Bildkarten.

3.4 Ende-zu-Ende-Extraktion mittels multimodaler Large Language Models

[22] Im Gegensatz zu einer wie oben beschriebenen, aus einzelnen Erkennungsschritten aufgebauten Verarbeitungspipeline stehen Ende-zu-Ende-Modelle, welche die geforderte Extraktion in einem Schritt übernehmen können. Beispielsweise zeigt Gewook Kim (2022) einen Ende-zu-Ende-Ansatz für die Extraktion von Schlüsselinformationen aus Kassenbelegen auf einem frei verfügbaren Standard-Datensatz.[8] Diese Ansätze erfordern jedoch eine große Menge an passendem Trainingsmaterial. Da die von uns untersuchten Dokumente in Deutsch verfasst sind und ungewöhnliche Schriftarten (Schreibmaschine) enthalten, sind solche Lösungen nur bedingt auf unsere Daten anwendbar, zumindest solange nicht größere Mengen an annotierten Daten für ein Nachtraining der Modelle verfügbar sind.

[23] Als mögliche Lösung zeichneten sich insbesondere multimodale Large Language Modelle für das Lösen vieler bis dahin nur mit Spezialalgorithmen anzugehender Probleme aus. Aus unseren Experimenten ging hervor, dass insbesondere GPT-4o[9] für die Erkennung von strukturierten Informationen in Dokumenten gute Ergebnisse erzielen kann. Die Eingabe besteht aus dem vorverarbeiteten Bild sowie einem Prompt (s. Abb. 3). Dieser beschreibt im Detail die erwarteten Kategorie-Wörter und definiert die Aufgabe, ausschließlich Texte (Einträge) aus dem Dokument wiederzugeben und diese den korrekten Kategorie-Wörtern zuzuordnen. Im Gegensatz zum ›klassischen Verfahren‹ besteht die Ausgabe von GPT-4o nur aus Text. In der Ausgabe sind folglich keine Metainformationen zur Position der Textelemente enthalten, jedoch ordnet das Modell gemäß dem Prompt die erkannten Einträge automatisch den entsprechenden Kategorien zu.

```
system_message = "Deine Aufgabe ist es strukturiert Daten aus einem Dokument zu extrahieren.\n"
                 "Die zu erkennenden Felder sind:\n"
                 "FO, Fdst, Land, Kreis, Mus, RGZM, Neg, Ggst, Zeit, Lit. und Photo\n"
                 "Genauere Beschreibung der Felder:\n"
                 "- FO: Der Fundort des Objektes\n"
                 "- Fdst: Die Fundstelle des Objektes\n"
                 "- Land: Das Bundesland oder Land, wo das Objekt gefunden / das Foto aufgenommen wurde\n"
                 "- Kreis: Der Landkreis, wo das Objekt gefunden wurde\n"
                 "- Mus: Das Museum\n"
                 "- RGZM: Die RGZM Nummer\n"
                 "- Neg: Negativnummer, oder Beschreibung wer das Negativ angefertigt hat\n"
                 "- Ggst: Beschreibung des Gegenstandes / bzw. Beschreibung der Abbildung\n"
                 "- Zeit\n"
                 "- Lit\n"
                 "- Photo\n"
                 "Erfinde nichts dazu, halte die an die Definition und die Information aus dem Dokumentenscan\n"
                 "Gebe die Schlüssel-Wert-Paare ohne Formatierung als Plaintext zurück.\n"
                 "Ist ein Feld leer oder im Original nicht vorhanden, lasse es auch in der Ausgabe leer\n"
user_message =   "Extrahiere die struktuierten Daten aus diesem Dokument: "
```

Abb. 3: Verwendeter Prompt für GPT-4o.

[6] Vgl. Shi 2016.
[7] Vgl. Kammleitner 2024.
[8] Vgl. Kim 2022.
[9] Vgl. OpenAI 2025; genaue Bezeichnung des Modells: gpt-4o-2024-08-06.

Die Zuordnung der Einträge zu den Kategorien in der Ausgabe des *LLMs* ist hierbei für den Menschen intuitiv einsehbar, für die Weiterverarbeitung in einer Datenbank müssen diese entsprechenden Kategorie-Wort-Text-Paare jedoch systematisch extrahiert werden (z. B. als strukturierte *JSON* -Objekte). Zu diesem Zweck verwendeten wir nicht das LLM, da dies in Experimenten zu leichten Qualitätseinbußen führte. Stattdessen wurden für jedes Kategorie-Wort mehrere reguläre Grammatiken definiert, die auf die Ausgabe von GPT-4o angewandt werden. Diese Grammatiken berücksichtigen verschiedene Schreibweisen der Kategorie-Wörter, sowohl in ausgeschriebener als auch in abgekürzter Form, sowie Platzhalter für potenzielle Sonder– und Trennzeichen. Darüber hinaus enthalten sie Mechanismen zur Identifikation und Extraktion des relevanten Wertes für jedes Kategorie-Wort. Die regulären Ausdrücke werden dabei für jedes Kategorie-Wort in einer vordefinierten Reihenfolge entsprechend der Schemavorgabe und dem Prompt auf die Ausgabe angewandt. Überflüssige Texte (wie beispielsweise ein Antwortsatz) werden dabei automatisch ignoriert, da diese nicht von den definierten Grammatiken abgedeckt werden. [24]

3.5 Wissensbasierte Nachbearbeitung

Es hat sich herausgestellt, dass die automatischen Algorithmen einige typische Fehler machen, die durch eine Nachbearbeitung teilweise korrigiert werden können. Das gilt sowohl für die Extraktions-Pipeline als auch für die LLMs. Wie bereits erwähnt, wurden längere Texte zu einer Kategorie häufig in freie Felder der darunterliegenden Kategorie weitergeschrieben, was zu Fehlern bei der Zuordnung führt. Um dies zu erkennen, wurden zwei erweiterte Regeln verwendet, um die Genauigkeit bei den Negativ– und RGZM-Inventar-Nummer zu verbessern. Da diese eine exakt vorgegebene Struktur haben, werden reguläre Ausdrücke verwendet, um festzustellen, ob diese Nummern den entsprechenden Kategorien korrekt zugeordnet werden. Wurden die Nummern in anderen Kategorien erkannt, so werden diese verschoben. Insbesondere wird so auch Text, welcher nicht als gültige RGZM-Inventar-Nummer identifiziert werden konnte, zu dem darüberliegenden Eintrag ›Museum‹ verschoben. [25]

Eine zweite Nachbearbeitungstechnik besteht wie oben erwähnt im Aufbau einer Datenbank für die Namen von Lokalisationen (Fundort, Fundstelle, Kreis, Land), Museen, Gegenstands– und Zeitangaben. Für jede Bildkarte, die manuell nachkorrigiert wurde (s. nächster Abschnitt), wird die Datenbank um die entsprechenden Namen erweitert, wobei die Zuordnung des Namens zur Kategorie und auch die Häufigkeit des Vorkommens gespeichert wird. Wenn bei einer neuen Bildkarte ein unbekannter Name vorkommt, der sehr ähnlich zu einem Namen in der Datenbank ist (gemessen durch die *Levensthein-Distanz* , die die Anzahl von notwendigen Buchstabenkorrekturen zählt), wird diese automatisch ersetzt. Das kann zu veränderten Ersetzungen führen,[10] und verbessert im Durchschnitt die Erkennung. Zusätzlich können im Editor zur manuellen Nachkorrektur alle unbekannten Namen hervorgehoben werden, um die manuelle Nachkorrektur zu unterstützen. [26]

Teilweise ist bei der Kategorie ›Museum‹ neben dem Namen des Museums auch die Inventarnummer des Objektes dieses Museums angegeben. Dadurch können die Museumsnamen in der Datenbank nicht wiedererkannt werden. Hier wäre noch eine weitere Nachbearbeitung nützlich, die beim Inhalt der Kategorie ›Museum‹ den Eintrag in den Namen und eine optionale Inventarnummer zerlegt; diese ist aber noch nicht implementiert. [27]

[10] Die Änderungen im Vergleich zum Original sind meist Vereinheitlichungen in der Schreibweise, z. B. wenn auf den Karteikarten teils »Großgartach« und teils »Grossgartach« steht, wird das einheitlich durch den Namen in der Datenbank ersetzt (»Großgartach«). In ähnlicher Weise wurde beim Fundort »Heilbronn« durch »Kr. Heilbronn« ersetzt, weil letzteres die häufigere, in der Datenbank gespeicherte Schreibweise ist.

3.6 Editor zur manuelle Nachkorrektur

Beide Ansätze erzeugen eine strukturierte Ausgabe im *JSON* -Format als Ergebnis. Dieses kann im Anschluss mit einem von den Autoren speziell für strukturierte Informationen entwickelten Korrekturtool verbessert werden, bei dem man die Kategorien, die erkannt werden sollen, mittels eines Schemas flexibel konfigurieren kann. Das Tool ist als Webanwendung konzipiert, ist aber auch vollständig als Desktop-Applikation lauffähig (Abb. 4). Die Anwendung stellt jede Bildkarte einzeln links im Original dar und auf der rechten Seite im Formular werden die extrahierten Einträge gemäß dem definierten Schema angezeigt. Diese können dann manuell auf Konsistenz geprüft und ggf. angepasst bzw. gelöscht werden und fehlende Einträge hinzugefügt. [28]

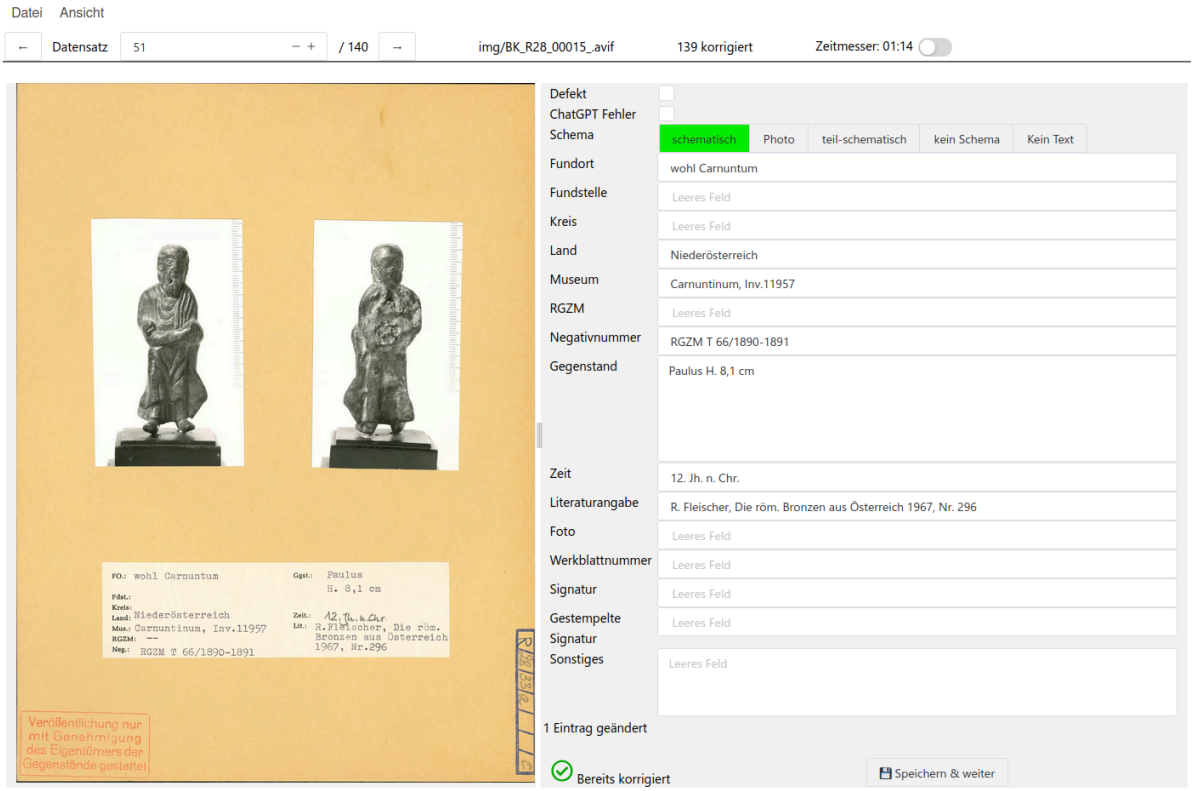

Abb. 4: Korrekturtool für die automatisch generierte Transkription der Bildkarten. Links die originale Bildkarte, rechts das Ergebnis der Transkription. Die letzten vier Kategorien (Foto, Werkblattnummer, Signatur, Gestempelte Signatur) wurden in den Evaluationen in Tab. 1 bis 4 nicht genutzt. In diesem Beispiel musste eine Korrektur vorgenommen werden, da der Text «Fleischer« bei der Kategorie «Lit:« wegen der Überlappung nicht korrekt erkannt wurde. Der Zeitmesser oben rechts ermöglichte die Zeitnahme für die Dauer der händischen Korrekturen. [Fotos: Leibniz-Zentrum für Archäologie (LEIZA) / LEIZA-Archiv]

Da das Verschieben von Texten (›Copy-und-Paste‹) sehr viel weniger Zeit erfordert als das Neuschreiben von Texten, wurde zusätzlich zum Transkriptionsergebnis mit schematischer Zuordnung noch ein zweites Transkriptionsergebnis des reinen Textes ohne Strukturierung angezeigt (mit einem anderen, sehr einfachen Prompt),[11] da bei der strukturierten Transkription manche Texte aus nicht nachvollziehbaren Gründen schlichtweg fehlen und diese dann aus der unstrukturierten Transkription, in der sie meist vorhanden sind, mit Copy-und-Paste an die richtige Stelle verschoben werden können. [29]

[11] Prompt: »Transkribiere den gesamten Text im Dokument«.

4. Experimente und Evaluationen

4.1 Experiment-Design

[30] Zur Untersuchung der Transkriptionsleistung wurden für die vorgestellten Lösungsansätze folgende Experimente mit denselben Bildkarten durchgeführt. Aus der Gesamtmenge der rund 150.000 Bildkarten wurde ein zufälliges Subset von 217 schematischen Bildkarten (Subset 1) und 55 teil-schematischen Bildkarten (Subset 2) ausgewählt.

[31] Bei der Auswahl der Karten aus der Gesamtmenge wurde darauf geachtet, dass die Auswahl der für die Entnahme ausgewählten Schubladen einen Querschnitt der Bildkartensammlung abbildet und die Reihenfolge bei der Entnahme aus den Schubladen nicht verändert wurde.

[32] Die schematischen Bildkarten wurden mit dem Pipeline-Ansatz und mit GPT-4o (verschiedene Versionen seit November 2024) transkribiert und evaluiert. Die teil-schematischen Bildkarten wurden ausschließlich mit GPT-4o transkribiert und evaluiert. Anschließend wurden die Ergebnisse wissensbasiert nachbearbeitet (s. o.).

[33] Zur Erstellung der *Ground Truth* (d. h. der korrekten Transkriptionen, GT) wurde das im obigen Kapitel beschriebene Korrekturtool verwendet.

[34] Für die Evaluation zählen wir für jedes Dokument und jede Rubrik die Anzahl exakt mit den *Ground-Truth*-Annotationen übereinstimmenden Einträge (TP), zusätzlich erkannte Einträge, die nicht im Original vorhanden sind (FP), und fälschlicherweise als nicht vorhanden (leer) erkannte Einträge (FN). Einträge, die in der *Ground Truth* annotiert, aber von dem jeweiligen Verfahren falsch erkannt wurden, zählen wir gesondert als falsch erkannte Einträge (FP+FN), welche sowohl bei *Precision* als auch bei *Recall* als Fehler gezählt werden. Wir berechnen für jede zu erkennende Kategorie auf dieser Basis die folgenden Metriken:

- *Accuracy*, die für jeden Eintrag einschließlich der leeren Einträge angibt, ob der Eintrag korrekt ist
- *Precision* aus den vollständig korrekt erkannten Einträgen (TP) und der Anzahl der falsch erkannten Einträge (FP; dazu zählen ausgefüllte fehlerhafte Einträge sowie zusätzliche Einträge in Feldern, die auf der Bildkarte nicht ausgefüllt sind)
- *Recall*, welcher das Verhältnis vollständig korrekt erkannter Einträge (TP) und der Gesamtanzahl der zu findenden Einträge (TP+FN) beschreibt
- *F1-Score*, das gewichtete harmonische Mittel aus *Precision* und *Recall*

[35] Um die beiden Verfahren möglichst gut vergleichen zu können, werden die Ausgaben beider Ansätze vor der Evaluation normiert. Dabei werden alle im Deutschen unüblichen Umlaute wie Akzente entfernt. Zudem entfernen wir auch alle Sonder-, Satz– und Leerzeichen. Insbesondere letztere werden häufig ohne erkennbares Muster zwischen Zahlen, Abkürzungen, Negativ-, Inventar– und RGZM-Nummern oder bei Silbentrennung am Zeilenende von den Algorithmen eingefügt, was aufgrund der strengen Evaluationsvorschrift, dass erkannte Einträge exakt übereinstimmen müssen, ohne Vereinheitlichung zu zusätzlichen Fehlern führen würde, obwohl die Einträge aus Sicht eines menschlichen Gutachters korrekt erkannt wurden.

4.2 Evaluationsergebnisse

Die Korrektur der vom *LLM* transkribierten 217 schematischen Bildkarten hat ca. 55 Minuten gedauert, d. h. knapp 4 Bildkarten pro Minute. Tabelle 1 und 2 zeigen die Ergebnisse der Erkennung mittels GPT-4o pro Kategorie mit und ohne Nachbearbeitung und Tabelle 3 zeigt die entsprechenden Ergebnisse des Pipeline-Tools mit Nachbearbeitung. Für die 55 teil-schematischen Bildkarten zeigt Tabelle 4 das Ergebnis mit GPT-4o ohne Nachbearbeitung. [36]

	Alle Einträge (inkl. Leere)			Beschränkung auf nicht-leere Einträge							
Schlüsselwort	**Korrekt erkannt**	**Accuracy**	**# Richtig erkannte leere Felder (TN)**	**# Einträge**	**# Richtig erkannte Einträge (TP)**	**# Falsch erkannte Einträge (FP + FN)**	**# Zu viel erkannte Einträge (FP)**	**# Nicht erkannte Einträge (FN)**	**Precision TP / (TP + FP)**	**Recall TP / (TP + FN)**	**F1-Score**
Fundort	215	**99.1%**	0	216	215	1	1	0	**99.1%**	**99.5%**	**99.3%**
Fundstelle	214	**98.6%**	167	50	47	3	0	0	**94.0%**	**94.0%**	**94.0%**
Kreis	217	**100.0%**	170	47	47	0	0	0	**100.0%**	**100.0%**	**100.0%**
Land	213	**98.2%**	82	135	131	1	0	3	**99.2%**	**97.0%**	**98.1%**
Museum	194	**89.4%**	77	138	117	21	2	0	**83.6%**	**84.8%**	**84.2%**
RGZM-Nr.	209	**96.3%**	190	26	19	4	1	3	**79.2%**	**73.1%**	**76.0%**
Negativ-Nr.	203	**93.5%**	3	214	200	14	0	0	**93.5%**	**93.5%**	**93.5%**
Gegenstand	207	**95.4%**	0	217	207	10	0	0	**95.4%**	**95.4%**	**95.4%**
Zeit	213	**98.2%**	53	164	160	3	0	1	**98.2%**	**97.6%**	**97.9%**
Literaturangabe	215	**99.1%**	196	20	19	1	1	0	**90.5%**	**95.0%**	**92.7%**
Mittelwert		**96.77%**							**93.25%**	**92.98%**	**93.10%**

Tab. 1: Evaluationsergebnisse von ChatGPT 4o für 217 schematische Bildkarten mit Nachbearbeitung.

	Alle Einträge (inkl. Leere)			Beschränkung auf nicht-leere Einträge							
Schlüsselwort	**Korrekt erkannt**	**Accuracy**	**# Richtig erkannte leere Felder (TN)**	**# Einträge**	**# Richtig erkannte Einträge (TP)**	**# Falsch erkannte Einträge (FP + FN)**	**# Zu viel erkannte Einträge (FP)**	**# Nicht erkannte Einträge (FN)**	**Precision TP / (TP + FP)**	**Recall TP / (TP + FN)**	**F1-Score**
Fundort	214	**98.6%**	1	216	213	3	0	0	**98.6%**	**98.6%**	**98.6%**
Fundstelle	216	**99.5%**	167	50	49	1	0	0	**98.0%**	**98.0%**	**98.0%**
Kreis	216	**99.5%**	170	47	46	0	0	1	**100.0%**	**97.9%**	**98.9%**
Land	215	**99.1%**	82	135	133	2	0	0	**98.5%**	**98.5%**	**98.5%**
Museum	176	**81.1%**	78	138	98	40	1	0	**70.5%**	**71.0%**	**70.8%**
RGZM-Nr.	177	**81.6%**	158	26	19	7	33	0	**32.2%**	**73.1%**	**44.7%**
Negativ-Nr.	200	**92.2%**	3	214	197	15	0	2	**92.9%**	**92.1%**	**92.5%**
Gegenstand	207	**95.4%**	0	217	207	10	0	0	**95.4%**	**95.4%**	**95.4%**
Zeit	212	**97.7%**	53	164	159	3	0	2	**98.1%**	**97.0%**	**97.5%**
Literaturangabe	214	**98.6%**	195	20	19	1	2	0	**86.4%**	**95.0%**	**90.5%**
Mittelwert		**94.33%**							**87.07%**	**91.65%**	**88.54%**

Tab. 2: Evaluationsergebnisse von GPT-4o für 217 schematische Bildkarten ohne Nachbearbeitung.

	Alle Einträge (inkl. Leere)			Beschränkung auf nicht-leere Einträge							
Schlüsselwort	**Korrekt erkannt**	**Accuracy**	**# Richtig erkannte leere Felder (TN)**	**# Einträge**	**# Richtig erkannte Einträge (TP)**	**# Falsch erkannte Einträge (FP + FN)**	**# Zu viel erkannte Einträge (FP)**	**# Nicht erkannte Einträge (FN)**	**Precision TP / (TP + FP)**	**Recall TP / (TP + FN)**	**F1-Score**
Fundort	194	**89.4%**	0	216	194	18	1	4	**91.1%**	**89.8%**	**90.4%**
Fundstelle	210	**96.8%**	165	50	45	4	2	1	**88.2%**	**90.0%**	**89.1%**
Kreis	215	**99.1%**	169	47	46	1	1	0	**95.8%**	**97.9%**	**96.8%**
Land	195	**89.9%**	79	135	116	15	3	4	**86.6%**	**85.9%**	**86.2%**
Museum	181	**83.4%**	77	138	104	34	2	0	**74.3%**	**75.4%**	**74.8%**
RGZM-Nr.	204	**94.0%**	191	26	13	7	0	6	**65.0%**	**50.0%**	**56.5%**
Negativ-Nr.	198	**91.2%**	3	214	195	13	0	6	**93.8%**	**91.1%**	**92.4%**
Gegenstand	186	**85.7%**	0	217	186	29	0	2	**86.5%**	**85.7%**	**86.1%**
Zeit	194	**89.4%**	52	164	142	6	1	16	**95.3%**	**86.6%**	**90.7%**
Literaturangabe	204	**94.0%**	197	20	7	13	0	0	**35.0%**	**35.0%**	**35.0%**
Mittelwert		**91.29%**							**81.16%**	**78.74%**	**79.82%**

Tab. 3: Evaluationsergebnisse des Pipeline-Tools für 217 schematische Bildkarten.

	Alle Einträge (inkl. Leere)			Beschränkung auf nicht-leere Einträge							
Schlüsselwort	Korrekt erkannt	Accuracy	# Richtig erkannte leere Felder (TN)	# Einträge	# Richtig erkannte Einträge (TP)	# Falsch erkannte Einträge (FP + FN)	# Zu viel erkannte Einträge (FP)	# Nicht erkannte Einträge (FN)	Precision TP / (TP + FP)	Recall TP / (TP + FN)	F1-Score
Lokalisation	45	**81.8**%	0	55	45	10	0	0	**81.8**%	**81.8**%	**81.8**%
Museum	41	**74.5**%	13	42	28	14	0	0	**66.7**%	**66.7**%	**66.7**%
RGZM-Nr.	48	**87.3**%	43	8	5	2	4	1	**45.5**%	**62.5**%	**52.6**%
Negativ-Nr.	39	**70.9**%	15	37	24	11	3	2	**63.2**%	**64.9**%	**64.0**%
Gegenstand	35	**63.6**%	7	45	28	15	3	2	**60.9**%	**62.2**%	**61.5**%
Zeit	50	**90.9**%	41	9	9	0	5	0	**64.3**%	**100.0**%	**78.3**%
Literaturangabe	54	**98.2**%	26	29	28	0	0	1	**100.0**%	**96.6**%	**98.2**%
Mittelwert		**81.04**%							**68.89**%	**76.37**%	**71.88**%

Tab. 4: Evaluationsergebnisse von GPT-4o für 55 teil-schematische Bildkarten. Die vier Kategorien Fundort, Fundstelle, Kreis und Land des Stempels der schematischen Bildkarten wurden zu einer Kategorie »Lokalisation« zusammengefasst, da bei den teil-schematischen Bildkarten deren Feinaufteilung nicht immer eindeutig ist.

Weitere Experimente umfassten die Erkennung der Signaturen mit GPT-4o am Rand der schematischen Bildkarten, die oftmals angegeben sind (vgl. Abb. 2). Die gestempelten Signaturnummern enthalten meist sechs gedruckte Ziffern mit oder ohne Schrägstriche wie z. B. »05/28/10« oder nur »052810«. Teilweise werden die Schrägstriche mit Ziffern, insbesondere der »1« verwechselt, so dass statt »05/28/10« z. B. »05128110« erkannt wird. [37]

In einer Stichprobe mit 45 gestempelten Signatureinträgen wurden 43 korrekt erkannt (95,5 %), wobei die gelegentlichen handschriftlichen Buchstaben hinter der Signatur ignoriert wurden. Die handschriftlichen Signaturen in den blauen Kästchen wurden dagegen deutlich schlechter erkannt. In einer Stichprobe mit 40 handschriftlichen Signaturen wurde 16 korrekt erkannt (40 %). Im Vergleich zu den sonstigen Kategorien ist deren Nachkorrektur relativ zeitaufwändig. Die Werkblatt-Nummern wurden im Experiment ignoriert, da sie keine relevanten Informationen zu den Objekten enthalten. [38]

4.3 Ergebnisse und Diskussion

Während der Pipeline-Ansatz bei den schematischen Bildkarten auf einen *F1-Score* von ca. 80 % kommt, liegt der *F1-Score* von GPT-4o ohne Nachbearbeitung bei 88,5 % und mit Nachbearbeitung bei ca. 93 % pro Kategorie, also deutlich höher. Bei beiden Methoden liegen die Hauptprobleme wie bereits angedeutet layoutbedingt in den Einträgen der Kategorien ›Museum‹ und ›RGZM‹: Der Text von ›Museum‹ passte nicht in die vorgesehene Zeile und wurde in der nächsten Zeile im Feld ›RGZM‹ fortgesetzt, was in 65 der 217 schematischen Karten erfolgte. In 33 dieser 65 Fälle wurden die Texte von GPT-4o nicht richtig erkannt oder zugeordnet, was sowohl bei der Kategorie ›Museum‹ als auch ›RGZM‹ als Fehler gezählt wurde (s. Tab. 2). Durch die Nachbearbeitung wurde in 32 der 33 Fälle der Text des Feldes ›RGZM‹ in das Feld von ›Museum‹ geschoben, was die Anzahl der Fehler deutlich reduziert, aber nicht vollständig eliminiert hat, da auch andere Fehler im Feld ›Museum‹ vorkommen (s. Tab. 1). [39]

Das Pipeline-Tool schneidet bei regulären Bildkarten sehr gut ab, aber es steht bei handschriftlichen Einträgen oder Überlagerungen von Stempel und inhaltlichem Text vor praktisch unlösbaren Problemen, da es darauf nicht trainiert wurde. Hier besteht grundsätzlich noch erhebliches Potenzial zur Verbesserung. Bezüglich GPT-4o wurde bisher mit einem relativ einfachen Prompt gearbeitet. Auch hier besteht Verbesserungspotenzial, wenn man in den Prompt Beispiele von schwierigen Bildkarten und deren korrekte Transkription hinzufügen würde. Das größte Verbesserungspotenzial liegt jedoch im Ausbau der Datenbank für die wissensbasierte Nachbearbeitung. Bisher wurden nur relativ wenige Bildkarten bearbeitet, so dass die Datenbank klein ist. Wenn das Tool für eine Massentranskription der über 100.000 Bildkarten benutzt wird, ist zu erwarten, dass die meisten Namen sich wiederholen und dadurch die Erkennungsrate sehr gute Ergebnisse zeigt. Das gilt auch für die teil-schematischen Bildkarten, die bisher deutlich niedrigere Erkennungsraten als die schematischen Bildkarten haben. [40]

Außerdem sollen die Signaturen und Werkstattnummern am rechten oder oberen Rand heuristisch nachbearbeitet werden, indem insbesondere überprüft wird, ob eine »1«, die häufig mit einem Schrägstrich »/« verwechselt wird, plausibel ist. Diese Plausibilitätsprüfung soll die Bedeutung der kodierten Signaturen mitberücksichtigen, die zusätzlich durch den Vergleich mit den erkannten Inhalten in den Kategorien der Karten validiert werden können. Die Signaturen sind zwar nicht auf allen Karten einheitlich, aber meistens werden in den ersten beiden Doppelziffern Zeitstellungen und Fundregionen der Motive kodiert. Schließlich soll diese heuristische Validierung auch für die Erkennung der – bislang schwer erkennbaren – handgeschriebenen Ziffern und Buchstaben genutzt werden. Weiterhin ist es vielversprechend, für die Erkennung der handgeschriebenen Ziffern spezielle *HTR*-Modelle zu trainieren, da es unklar ist, wie schnell sich *LLMs* für diesen Anwendungsfall adaptieren und verbessern. Dies ist vielversprechend, da die Ziffern in gut lesbarer Druckschrift eingetragen sind und wir vermuten, dass für die Texterkennung die größte Hürde die gestempelte Schablone darstellt. Wir gehen davon aus, dass durch ein Nachtraining die Modelle auch lernen, die Schablone korrekt zu interpretieren. [41]

5. Zusammenfassung und Ausblick

Die Ergebnisse der Experimente zeigen, dass mit den jetzigen Tools bereits eine große Effizienzsteigerung im Vergleich zu einer manuellen Transkription der Bildkarten erzielt werden kann, da ein Großteil der zu erkennenden Informationen bereits korrekt erkannt wurde. Im Vergleich zeigt sich GPT-4o dem Pipeline-Ansatz deutlich überlegen. Dies lässt sich insbesondere auf die robustere Erkennung schwieriger Fälle zurückführen, beispielsweise, wenn einzelne Einträge handschriftlich verfasst sind, Text sehr dicht beieinander oder überlagernd gedruckt ist oder der Text durch einen Stempel überlagert wird. [42]

Die manuelle Nachbearbeitung auf Basis der algorithmischen Transkriptionen und Zuordnungen ist bereits jetzt sehr effizient und kann durch Hervorheben unsicherer Transkriptionen im Korrekturtool noch effizienter gemacht werden. Der damit verbundene Aufbau einer großen Datenbank bei einer Massentranskription lässt weitere Effizienzgewinne erwarten. [43]

Eine größere Herausforderung, die wir bisher nicht bearbeitet haben, ist die Transkription der nicht-schematischen Bildkarten, da diese weit heterogener sind als die schematischen und teil-schematischen Bildkarten. Damit auch diese strukturiert transkribiert werden können, ist zunächst eine Einteilung in Untergruppen mit Karten ähnlicher Struktur beabsichtigt. Wir erwarten, dass auch deren Strukturen dann, möglicherweise unter Zuhilfenahme multimodaler *LLMs*, automatisch erkannt und die erkannten Kategorien dann für jeden Kartentyp bearbeitet werden können. [44]

Über die archivalische Erschließung hinaus können die auf diese Weise effizient extrahierten und strukturierten Textinformationen sehr vieler Bildkarten auf mehrfache Weise genutzt werden: Neben der semantischen Suche und datenbasierten Auswertungen im Sinne der Digital Humanities können diese auch zur automatisierten Klassifikation und Beschreibung von Abbildungen ähnlicher historischer oder archäologischer Motive genutzt werden, zu denen keine oder weniger detaillierte Textinformationen vorliegen oder die noch nicht anders erschlossen worden sind. Die Grundidee besteht darin, die beschreibenden Texte zur Annotation der Bilder zu verwenden und dann aus strukturiert annotierten Bilddaten deren Klassifikation nach verschiedenen Aspekten mit *Deep-Learning*-Methoden zu lernen. Der gezeigte Ansatz bietet somit vielfältige Perspektiven zur Erschließung und weiteren Nutzung von Archivalien mit Text- und Bild-Informationen. [45]

Bibliografie

Taylor Arnold / Lauren Tilton: Distant Viewing. Cambridge, US-MA 2023. PDF. DOI: 10.7551/mitpress/14046.001.0001

Norbert Fischer / Alexander Hartelt / Frank Puppe: Line-Level Layout Recognition of Historical Documents with Background Knowledge. In: Algorithms 16 (2023), H. 3. S. 136. 2023. PDF. DOI: 10.3390/a16030136

Herder-Institut für historische Ostmitteleuropaforschung (Hg.): Conference: Artifical Intelligence in Archives and Collections. Letzter Zugriff: 14.07.2025. HTML. [online]

Martin Kammleiter: Extraktion strukturierter Daten aus Museums-Bildkarton-Karten mittels Deep-Learning und wissensbasierten Methoden. Bachelorarbeit (betreut von Frank Puppe und Norbert Fischer). Universität Würzburg 2024.

Geewook Kim / Teakgyu Hong / Moonbin Yim / JeongYeon Nam / Jinyoung Park / Jinyeong Yim / Wonseok Hwang / Sangdoo Yun / Dongyoon Han / Seunghyun Park: OCR-Free Document Understanding Transformer. In: Shai Avidan / Gabriel Brostow / Moustapha Cissé / Giovanni Maria Farinella / Tal Hassner (Hrsg.): Computer Vision – ECCV 2022 (= Lecture Notes in Computer Science, 13688), S. 498–517. Cham 2022. PDF. DOI: 10.1007/978-3-031-19815-1_29

Leibniz-Zentrum für Zeithistorische Forschung Potsdam (Hg.): Value of the Past. Letzter Zugriff: 23.02.2025. HTML. [online]

Haofu Liao / Aruni RoyChowdhury / Weijian Li / Ankan Bansal / Yuting Zhang / Zhuowen Tu / Ravi Kumar Satzoda / R. Manmatha / Vijay Mahadevan: DocTr: Document Transformer for Structured Information Extraction in Documents. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris 2023. S. 19527-19537. PDF. 10.1109/ICCV51070.2023.01794

OpenAI: Hello GPT-4o. Veröffentlichungszeitpunkt: 03.2023. Letzter Zugriff: 23.02.2025. HTML. [online]

Baoguang Shi / Xiang Bai / Cong Yao: An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2017), H. 11. S. 2298-2304. PDF. DOI: 10.1109/TPAMI.2016.2646371

Thomas Smits / Melvin Wevers: A Multimodal Turn in Digital Humanities: Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections. In: *Digital Scholarship in the Humanities* 38 (2023), S. 1267-1280. PDF. DOI: 10.1093/llc/fqad008

Melvin Wevers / Thomas Smits: The Visual Digital Turn: Using Neural Networks to Study Historical Images. In: Digital Scholarship in the Humanities 35 (2020), H. 1. S. 194-207. PDF. DOI: 10.1093/llc/fqy085

Abbildungsverzeichnis

Abb. 1: Verschiedene Schemata der Bildkarten im LEIZA-Archiv. Oben: Zwei schematisch beschriebene Bildkarten basierend auf einem – teilweise vorgedruckten – ›Formular‹, dessen Einträge mit Schreibmaschine ausgefüllt wurden (links: einfaches Layout, rechts: kompliziertes Layout); unten: zwei teil-schematisch beschriebene Bildkarten, bei denen nur manche Kategorie-Namen des obigen ›Formulars‹ angegeben sind. [Fotos: unten links: Simone Deyts / Claude Rolley: L'Art de la Bourgogne romaine, découvertes récentes. Musée archéologique de Dijon, France 1973, Kat. Nr. 25; unten rechts: Mainzer Zeitschrift 36, 1941 Taf. II,1; alle LEIZA-Archiv]

Abb. 2: Signaturen auf Bildkarten des LEIZA-Bildarchivs. Beispiele für drei gestempelte Signaturen (links), eine handschriftliche Werkblatt-Nummer unter dem nur rudimentär ausgefüllten Signaturfeld (Mitte) und eine handschriftliche Signatur in einem gestempelten blauen Kästchen (rechts). [Ausschnitte aus Bildkarten, LEIZA-Archiv].

Abb. 3: Verwendeter Prompt für GPT-4o.

Abb. 4: Korrekturtool für die automatisch generierte Transkription der Bildkarten. Links die originale Bildkarte, rechts das Ergebnis der Transkription. Die letzten vier Kategorien (Foto, Werkblattnummer, Signatur, Gestempelte Signatur) wurden in den Evaluationen in Tab. 1 bis 4 nicht genutzt. In diesem Beispiel musste eine Korrektur vorgenommen werden, da der Text «Fleischer« bei der Kategorie «Lit:« wegen der Überlappung nicht korrekt erkannt wurde. Der Zeitmesser oben rechts ermöglichte die Zeitnahme für die Dauer der händischen Korrekturen. [Fotos: Leibniz-Zentrum für Archäologie (LEIZA) / LEIZA-Archiv]

Tab. 1: Evaluationsergebnisse von ChatGPT 4o für 217 schematische Bildkarten mit Nachbearbeitung.

Tab. 2: Evaluationsergebnisse von GPT-4o für 217 schematische Bildkarten ohne Nachbearbeitung.

Tab. 3: Evaluationsergebnisse des Pipeline-Tools für 217 schematische Bildkarten.

Tab. 4: Evaluationsergebnisse von GPT-4o für 55 teil-schematische Bildkarten. Die vier Kategorien Fundort, Fundstelle, Kreis und Land des Stempels der schematischen Bildkarten wurden zu einer Kategorie »Lokalisation« zusammengefasst, da bei den teil-schematischen Bildkarten deren Feinaufteilung nicht immer eindeutig ist.