Künstliche Intelligenz entziffert historische Handschriften

Blick in ein Magazin des Landesarchivs Sachsen-Anhalt

Längst ist Künstliche Intelligenz (KI) in der Gegenwart angekommen und kein Thema mehr, das allein Filmen und Romanen der Science-Fiction vorbehalten ist. Im Rahmen einer Transferarbeit testete nun auch das Landesarchiv Sachsen-Anhalt den Einsatz von KI.

Archive als einzigartige Schatzkammern des Wissens

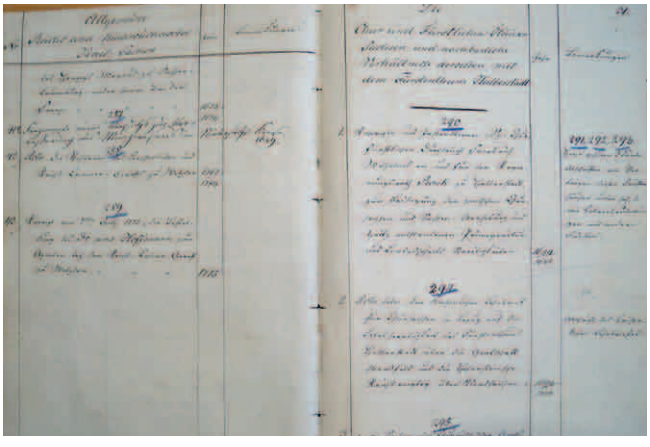
An keinem Ort der Welt lässt sich mehr über die Geschichte Sachsen-Anhalts und seiner Vorgängerterritorien erfahren als im zuständigen Landesarchiv. Was unscheinbar als Bündel oder verpackt in Mappen und Schachteln in den langen Regalreihen der wohlgeordneten Magazine lagert, sind einzigartige Kostbarkeiten. Auf mehr als 64 Regalkilometern finden sich historische Unterlagen aus mehr als 1.000 Jahren, die Generationen von Menschen hinterlassen haben. Noch immer warten unzählige Geheimnisse in den Archiven darauf, entdeckt zu werden.

Unvollständige Wegweiser zu historischen Dokumenten

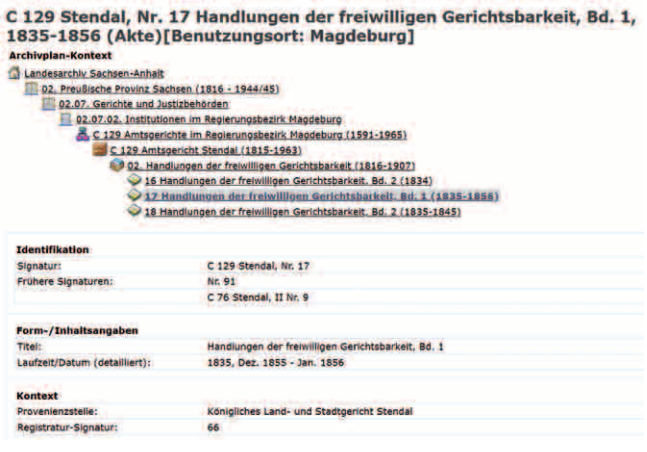
Zwar steht dieser reiche Schatz an historischem Wissen grundsätzlich jeder Person zur Verfügung, doch stellen sich auf dem Weg dorthin zweierlei Arten von Hürden. So bedarf es einerseits eines Wegweisers, um unter den vielen Millionen Archivalien die gewünschten Informationen zu finden.

Als Grundlage dafür erfasst das Archivpersonal in möglichst knapper Form wesentliche Grundinformationen der Unterlagen. Die Ergebnisse gingen einst in handgeschriebene und später maschinenschriftliche Findbücher ein und werden mittlerweile in digitalen Archivinformationssystemen veröffentlicht. Angesichts sinkender personeller Kapazitäten sowie der erheblichen Masse an Dokumenten, die das Archivpersonal zu bearbeiten hat, kann dieser Wegweiser allerdings nicht tagesaktuell sein. Vielmehr bestehen im Landesarchiv ähnlich wie in anderen Archiven erhebliche Erschließungsrückstände. So kam etwa das Bundesarchiv im Jahr 2017 zu der ernüchternden Erkenntnis, dass es 2.400 Personenjahre benötigte, um alle Rückstände in der Erschließung aufzuarbeiten. Folglich existiert mit den Findbüchern und dem Archivinformationssystem nur ein unvollständiger Wegweiser, oder anders ausgedrückt, ein Routenplaner mit veraltetem Update, um sich in den Inhalten der Archivalien orientieren zu können.

Handschriftliches Findbuch aus dem 19. Jahrhundert



Rechercheergebnis im Archivinformationssystem scopeArchiv



C 129 Stendal, Nr. 17 Handlungen der freiwilligen Gerichtsbarkeit, Bd. 1, 1835-1856 (Akte)[Benutzungsort: Magdeburg]

Archivplan-Kontext

Landesarchiv Sachsen-Anhalt
02. Preußische Provinz Sachsen (1816 - 1944/45)
02.07. Gerichte und Justizbehörden
02.07.02. Institutionen im Regierungsbezirk Magdeburg
C 129 Amtsgerichte im Regierungsbezirk Magdeburg (1591-1965)
C 129 Amtsgericht Stendal (1815-1963)
02. Handlungen der freiwilligen Gerichtsbarkeit (1816-1907)
16 Handlungen der freiwilligen Gerichtsbarkeit, Bd. 2 (1834)
17 Handlungen der freiwilligen Gerichtsbarkeit, Bd. 1 (1835-1856)
18 Handlungen der freiwilligen Gerichtsbarkeit, Bd. 2 (1835-1845)

Identifikation

Signatur:	C 129 Stendal, Nr. 17
Frühere Signaturen:	Nr. 91 C 76 Stendal, II Nr. 9

Form-/Inhaltsangaben

Titel:	Handlungen der freiwilligen Gerichtsbarkeit, Bd. 1
Laufzeit/Datum (detailliert):	1835, Dez. 1855 - Jan. 1856

Kontext

Provenienzstelle:	Königliches Land- und Stadtgericht Stendal
Registratur-Signatur:	66

Handgemachte Schrifträtsel

Doch selbst, wenn die Nutzenden das gewünschte Dokument gefunden und im Lesesaal oder am Bildschirm vor sich sehen, trennt sie häufig ein weiteres Problem von den Inhalten: die Handschrift. Während deren Lesbarkeit grundsätzlich vom Talent der schreibenden Person abhängt, kommt für die Zeit vor September 1941 der Gebrauch einer anderen Schrift hinzu. Denn im Unterschied zur heutigen lateinischen Ausgangsschrift verwendeten Schreibende bis dahin gewöhnlich die Deutsche Kurrentschrift als Standard im deutschen Sprachraum – mit teils gänzlich anders aussehenden Buchstaben.

Ob es zu den Aufgaben von Archiven gehört, dieses Schriftproblem der Nutzenden zu lösen, ist in der Fachwelt umstritten. Sofern sich Archive aber als Informationsdienstleister in der Wissensgesellschaft verstehen und die demokratische Nutzung ihrer Unterlagen erhöhen möchten, dürfte das Erarbeiten von Lösungsstrategien unumgänglich sein.

Automatische Handschriftenerkennung als Lösungsansatz

Im Rahmen seiner Transferarbeit diskutierte der Verfasser, ob Künstliche Intelligenz einen Beitrag zur Lösung dieser Probleme leisten kann. Ausgehend von der Beobachtung, dass eine KI-gestützte Erschließung zwingend maschinenlesbare Zeichen voraussetzt, widmete er sich schwerpunktmäßig der automatischen Handschriftenerkennung anhand der Software Transkribus. Leitend war die Frage, ob und unter welchen Voraussetzungen der Einsatz dieser Software im Landesarchiv sinnvoll erscheint.

Transkribus – eine genossenschaftliche Software

Finanziell unterstützt durch Projekte der Europäischen Union, entstand die Software Transkribus seit dem Jahr 2013. Übergeordnetes Ziel war es, historische Dokumente in eine virtuelle Forschungsumgebung einzubinden und sie digital durchsuchbar zu gestalten. Entsprechend des Produktnamens ist es die Kernaufgabe der Software, Texte zu transkribieren. In anderen Worten ausgedrückt, überträgt Transkribus also Buchstaben von einer Schrift in eine andere.

Nach dem Projektende im Jahr 2020 wechselte die Organisationsform unter der Bezeichnung READ co:op SCE zu einer europäischen Genossenschaft. Mit Stand von August 2023 gehörten dieser 135 Mitglieder aus 30 Ländern an. Abgesehen von der eigentlichen Texterkennung sind alle Funktionen der Software kostenlos verfügbar und lassen sich unter https://readcoop.eu/de/transkribus/ testen.

Titelblatt des Protokollbuchs (LASA, C 129 Stendal, Nr. 17)

Testobjekt Protokollbuch

Aus der Vielzahl von Archivalien, die in den Magazinen des Landesarchivs lagern, diente ein handgeschriebenes Protokollbuch als Gegenstand des praktischen Tests. Es ist unter der Signatur C 129 Stendal, Nr. 17 verzeichnet und als Digitalisat online unter http://recherche.landesarchiv.sachsen-anhalt.de/digital/C_129_Stendal__Nr_17.xml einzusehen.

Mehrere Schreiber notierten darin vor mehr als 150 Jahren die Amtshandlungen des Land- und Stadtgerichts Stendal, konkret für die Monate Januar bis Juni 1835 sowie Dezember 1855 bis Januar 1856. Auf 290 Blättern finden sich beispielsweise Nachweise über Bürgschaften und Testamente, ebenso Aufzeichnungen zu Pacht- und Kaufverträgen.

Warum Amtsbücher wie dieses als besonders reizvoller Gegenstand einer automatischen Handschriftenerkennung erscheinen, erklärt sich einerseits dadurch, dass sie nicht nur in den Magazinen des Landesarchivs massenhaft überliefert sind. Andererseits kommen sehr allgemeine Erschließungsinformationen hinzu: Zwar können Nutzende die Amtsbücher über das Archivinformationssystem finden, doch erfahren sie darin kaum etwas über die Inhalte. Vor diesem Hintergrund werden Amtsbücher bisher deutlich seltener eingesehen, als es ihre Bedeutung für verschiedene Gruppen von Nutzenden erwarten lässt.

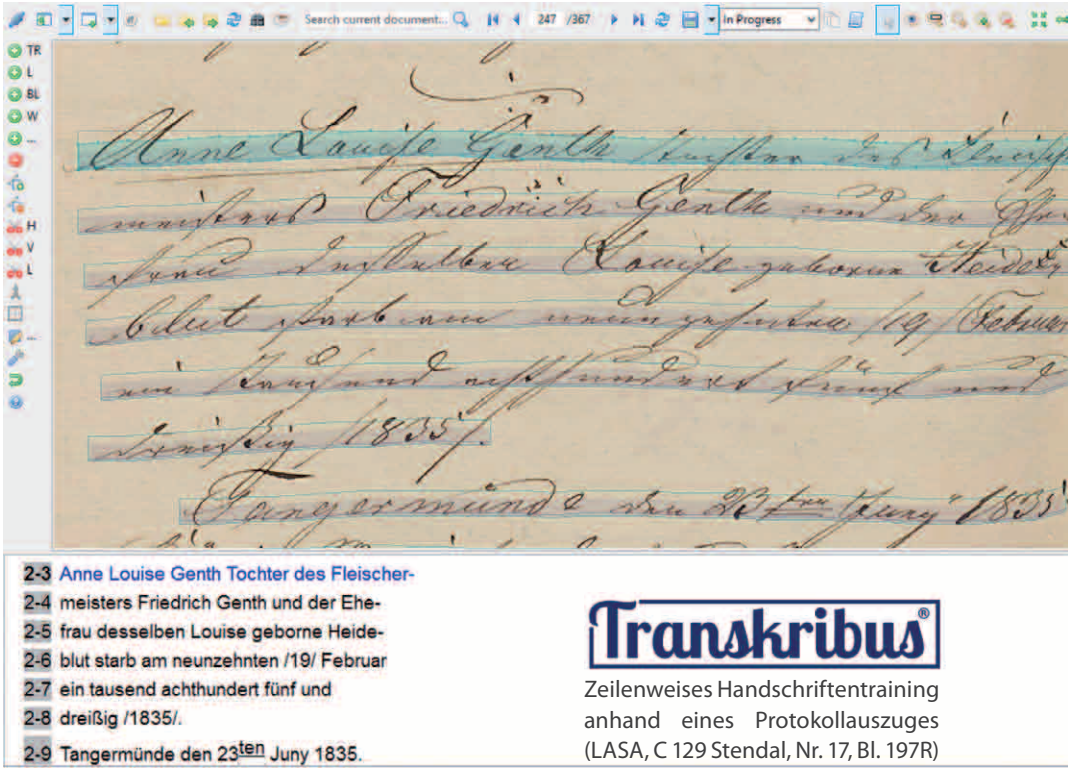
Vom Original zur maschinenlesbaren Transkription

Was Transkribus für seine Arbeit benötigt, sind digitale Abbilder der Originale. Diese werden in die Software geladen, von wo aus sich alle weiteren Schritte bis hin zur automatischen Handschriftenerkennung und ihrer Weiterverarbeitung durchführen lassen.

In technologischer Hinsicht bedient sich Transkribus der Handwritten Text Recognition (HTR). Dabei analysiert die Künstliche Intelligenz, stark vereinfacht ausgedrückt, die Textzeilen des digitalen Abbilds und vergleicht sie mit einer Datenbank, in der sich

eine Vielzahl von Schriftzeilen mitsamt Transkriptionen befindet. Je höher die Übereinstimmung zwischen dem digitalen Abbild und diesem so genannten Trainingsmodell ausfällt, desto fehlerfreier die automatisch zugeordnete Transkription.

Dementsprechend müssen in der Software zunächst die Textzeilen bestimmt werden. Hierzu bietet Transkribus einen Automatismus zur Layoutanalyse, der die Möglichkeit einer Nachbearbeitung umfasst.



Zeilenweises Handschriftentraining anhand eines Protokollauszuges (LASA, C 129 Stendal, Nr. 17, Bl. 197R)

Sobald die Software auf diese Weise gelernt hat, wo in den digitalen Abbildern Zeilen verlaufen, ließe sich bereits die automatische Handschriftenerkennung durchführen. Als Voraussetzung dafür benötigt Transkribus jedoch die erwähnte Datenbank mit Vergleichsmöglichkeiten. Diese lässt sich entweder selbst trainieren, indem Nutzende in der Software eine Auswahl von Seiten transkribieren. Alternativ besteht die Variante, öffentliche Trainingsmodelle zu verwenden, die ursprünglich für andere Dokumente entstanden sind.

Die eigentliche automatische Handschriftenerkennung erfordert nur wenige Einstellungen und ist per Mausklick auszulösen. Sobald das Ergebnis vorliegt, lässt sich die Qualität automatisch durch einen Vergleich mit idealen Transkriptionen bestimmen. Als Maß dienen die Zeichenfehlerrate (CER) und Wortfehlerrate (WER).

Zusätzlich bietet Transkribus weitere hilfreiche Funktionen, etwa das Kennzeichnen von besonderen Inhalten wie Orts- und Personennamen oder die Präsentation der Transkriptionsergebnisse auf einer read&search-Homepage. Letztere überzeugt zudem durch ihre innovative Suchfunktion.

Erfahrungen aus der Praxis

Grundsätzlich erwies sich der Einsatz von Transkribus am Beispiel des amtsgerichtlichen Protokollbuchs als komfortabel in der Handhabung. Zwar überschritt das Testergebnis eine Zeichenfehlerrate von 5 % beziehungsweise 25 Fehlern pro 500 Zeichen, wie sie die Deutsche Forschungsgemeinschaft in ihren Praxisregeln Digitalisierung als Höchstgrenze für akzeptable Ergebnisse der HTR-Technologie definiert. Die Ursache dafür lag aber weniger in der Software selbst als vielmehr an ihrer Verwendungsweise. Folglich ergaben sich wichtige Hinweise für mögliche künftige Projekte. Insbesondere die Layoutanalyse erwies sich als arbeits- und zeitaufwendig, indem der Automatismus die Zeilen trotz verschiedener Einstellungen allein in etwas mehr als drei Vierteln der Digitalisate fehlerfrei markierte. Häufige Fehlerquellen lagen einerseits im Bereich der korrekten Abgrenzung von Textregionen, da die Software keine Trennung zwischen nahe beieinanderliegenden Zeilen umsetzte. Andererseits blieben im Bereich der Vollständigkeit erfasster Textzeilen feine sowie abseits der Haupttextblöcke stehende Zeichen wiederholt unberücksichtigt.

Aufgrund der Vielgestaltigkeit seiner beschriebenen Seiten erwies sich das Protokollbuch als problematischer Gegenstand für eine automatische Handschriftenerkennung. Ein besonderer Einfluss auf das Verhältnis zwischen betriebswirtschaftlicher Effizienz und qualitativ hochwertigen HTR-Ergebnissen lässt sich mittels der Materialauswahl erzielen. Möglichst wenige unterschiedliche Handschriften in möglichst ähnlicher Seitengestaltung, digitalisiert als Einzelseiten und ohne erkennbare Zeichen darunterliegender Seiten, sind hierzu am ehesten geeignet. Bei vielgestaltigen Quellen wie dem Protokollbuch empfiehlt es sich, selbst ein Strukturmodell zu trainieren sowie möglichst alle enthaltenen Handschriften im empfohlenem Maß von etwa 100 bis 150 Seiten pro Handschrift zu berücksichtigen. Eine effiziente Verbesserung der Trainingsmodelle lässt sich bei ausreichender Qualität durch Korrektur automatisch erzeugter Transkriptionen erwirken. Manuelle Vor- und Nacharbeiten gehören dennoch zwingend zum Einsatz von Transkribus dazu, weshalb entsprechende Personalkapazitäten vorhanden sein müssen.

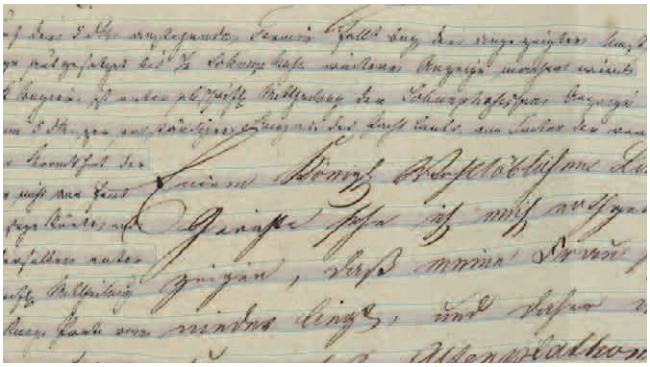
Konsequenzen für die Erschließung

Volltexttranskriptionen, wie sie Transkribus erzeugt, stellen nach vorherrschender Meinung keine eigenständige archivische Erschließung dar. Dennoch könnte ein Einsatz der Software die Erschließungspraxis unterstützen. So ermöglichen Volltexttranskriptionen etwa, dass Unterlagen auch von Personal ohne ausgeprägte Kenntnisse im Transkribieren historischer Handschriften erschlossen werden könnten. Außerdem bilden Transkriptionen in maschinenlesbaren Zeichen die Voraussetzung, um aus historischen Unterlagen automatisiert Inhalte zu entnehmen. Sofern Informationen derselben Kategorie stets in einheitlichen Bereichen der Seite erscheinen, ist es bereits gegenwärtig möglich, gewünschte Angaben massenhaft zu entnehmen und in die gewünschten Felder des Archivinformationssystems zu übertragen. Auf diese Weise hat etwa die Abteilung Osnabrück des Niedersächsischen Landesarchivs eine Kartei der Geheimen Staatspolizei erschlossen und die deutliche Ersparnis von Kosten und Personal gelobt. Perspektivisch steht zu erwarten, dass ähnliche Technologien auch für weniger einheitliche Quellenarten Einsatz finden werden.

Ansprüche der Nutzenden

Abgesehen davon lässt sich jedoch hinterfragen, ob die bisherige Erschließungspraxis noch den gegenwärtigen Ansprüchen der Nutzenden genügt. „Quod non est in Tela Totius Terrae, non est in mundo" – „was nicht im Internet zu finden ist, ist nicht in der Welt", ließe sich in Anlehnung an den bekannten Rechtsgrundsatz behaupten.

Transkribus böte Möglichkeiten, um den veränderten Suchgewohnheiten zu entsprechen. Neben Volltexttranskriptionen als betriebswirtschaftlich sinnvollem Ersatz für Tiefenerschließungen und dem Markieren von Normdaten bietet die Software mit dem Keyword Spotting eine innovative Suchfunktion: Anders als bei gewöhnlichen Volltextsuchen vergleicht die Software das Suchstichwort nicht mit festgelegten Zeichen, sondern ermittelt das Suchergebnis nach Wahrscheinlichkeiten der Übereinstimmung. Auf diese Weise lässt sich zudem das Hindernis einer uneinheitlichen Rechtschreibung ausgleichen, indem automatisch etwa verschiedene Schreibweisen eines Familien- oder Ortsnamens in den Suchergebnissen erscheinen. Beeindruckende Ergebnisse, wie sie etwa für die 256-bändigen Ratsprotokolle der Stadt Bautzen von 1623 bis 1832 (https://transkribus.eu/r/bautzen-ratsprotokolle) vorliegen, bestätigen den Nutzen der Technologie.



Nahe beieinanderliegende Zeilen als Fehlerquelle der Layoutanalyse (LASA, C 129 Stendal, Nr. 17, Bl. 176V)

Ausblick

Mit Blick in die Zukunft steht zu erwarten, dass Künstliche Intelligenz die archivarische Arbeit zunehmend beeinflussen wird. Wenn gegenwärtig bereits Software wie ChatGPT im Stande ist, wissenschaftliche Hausarbeiten zu verfassen, warum sollte perspektivisch nicht auch Erschließung vollkommen von KI übernommen werden können?

Fraglich bleibt, wie die Qualität der daraus resultierenden Erschließungsdaten gewahrt bleibt. Archivarisches Fachpersonal, so die Prognose, wird dazu auch künftig nicht obsolet werden. Überhaupt steht angesichts einer Stellungnahme des Deutschen Ethikrats vom 20. März 2023 zu diskutieren, in welchen Grenzen der Einsatz von KI auch in der archivarischen Arbeit als ethisch angemessen zu bewerten ist.

Diskutabel erscheint daneben die Frage nach den ökonomischen und ökologischen Folgen. Schließlich entstehen laufende Kosten und Energieverbräuche nicht allein durch das Herstellen und dauerhafte Bereitstellen von digitalen Reproduktionen ursprünglich analoger Unterlagen, sondern ebenso durch den Einsatz der Künstlichen Intelligenz.

Unabhängig davon gilt es, in den Archiven die Voraussetzungen für den Einsatz derartiger Zukunftstechnologien zu schaffen. Konkret bedarf es nicht allein fortwährender Digitalisierungen historischer Handschriften, sondern auch eines Engagements in der Bereitstellung maschinenlesbarer Volltexttranskriptionen. Wenn die Archive ihre Rolle als Informationsdienstleister wahren und erreichen möchten, dass die von ihnen überlieferten Daten einer großen Öffentlichkeit über zeitgemäße Recherchemöglichkeiten zugänglich werden, dürfen sie sich dieser Entwicklung nicht verschließen und müssen die digitale Transformation mitgestalten. Die automatische Handschriftenerkennung bietet das Potential, als wesentliches Mittel dafür zu dienen.

Marcel Giffey