Die Erstellung wissenschaftlicher Forschungsdaten aus analogen Beständen ist zu einer der Hauptaufgaben wissenschaftlicher Bibliotheken geworden. Die digitale Verfügbarkeit solcher Daten ist nicht zuletzt für die Disziplin der Digital Humanities unverzichtbar und bildet die Grundlage für computergestützte Analyseansätze in den Geistes- und Kulturwissenschaften. Während die Erschließung gedruckter Quellen voll automatisiert und nahezu fehlerfrei erfolgt, müssen handschriftliche Aufzeichnungen weiterhin manuell erschlossen werden. Dennoch führen Fortschritte in der *Optical Character Recognition* (OCR), wie der Einsatz neuronaler Netzwerke, die anhand von Trainingsdaten lernen, Zeichen zu entziffern, zu immer niedrigeren Fehlerraten. In der Erstellung von Trainingsdaten sowie in der Korrektur der automatisch erkannten Daten weisen digitale Transkriptionswerkzeuge jedoch noch erhebliche Probleme auf. Der Beitrag beschreibt die nutzerzentrierte Entwicklung eines Transkriptionstools zur Erschließung handschriftlicher Wetteraufzeichnungen aus dem 18. und 19. Jahrhundert. Mithilfe des Tools sollen die Aufzeichnungen computergestützt erschlossen und für geistes- und kulturwissenschaftliche Forschungszwecke zur Verfügung gestellt werden.

Producing academic research data from analogue holdings has become one of the principal tasks of academic libraries. The availability of such digital data is indispensable, not least for the discipline of Digital Humanities. It provides the foundation for computer-based analytical approaches in the humanities and cultural sciences. The cataloguing of printed sources is fully automated and virtually error-free, whereas handwritten records still need to be processed manually. Nevertheless, advances in *optical character recognition* (OCR), such as the use of neural networks that learn to decipher characters based on training data, are leading to ever lower error rates. However, digital transcription tools still have considerable difficulty in creating training data and correcting automatically recognised data. The paper describes the user-centred development of a transcription tool for the cataloguing of handwritten weather records from the 18th and 19th centuries. The tool will help in the computer-aided generation of the records and thus make them available for research purposes in the humanities and cultural sciences.

CONSTANTIN LEHENMEIER

Von historischen Wetteraufzeichnungen zu digitalen Forschungsdaten

Die Entwicklung eines Transkriptionstools mit Schwerpunkt Tabellen an der Universitätsbibliothek Regensburg (Teil I)

Besonders mit dem verstärkten Einsatz digitaler Werkzeuge in der geistes- und kulturwissenschaftlichen Forschung werden die vollständige Erschließung und Zugänglichmachung analoger Bestände zu einer Herausforderung, der sich wissenschaftliche Bibliotheken wie auch Archive und Museen stellen müssen. Historische Forschungsdaten sollen gewissenhaft aus gedruckten sowie handschriftlichen Quellen erzeugt und von Forschenden für systematische Analysen genutzt werden können. Im vorliegenden Beitrag werden anhand eines interdisziplinären Projekts der UB Regensburg sowohl Möglichkeiten und Probleme der computergestützten Erschließung als auch die damit einhergehenden Aufgabenbereiche von Bibliotheken, die über die bloße Lesbarmachung von Daten hinausgehen, aufgezeigt.

So werden im Zuge des *digital turn* und der Entmaterialisierung weitreichende Kompetenzen zur Digitalisierung, der standardisierten Aufbereitung und dem elektronischen Publizieren von Daten bis hin zur Schaffung von Schnittstellen gefordert.[1] Schließlich entsteht das digitale Potenzial analoger Bestände erst, wenn Dokumente und Texte in maschinenlesbarer, gesäuberter, mit strukturellem Markup ausgestatteter Form vorliegen und zugänglich sind.[2] Diese qualitativ hochwertige Tiefenerschließung und Zugänglichmachung von analogen Beständen bildet letztendlich die Grundlage für die wissenschaftliche Analyse und Interpretation mithilfe

digitaler Methoden wie automatisierter Bilderkennung, Text- und Data-Mining oder Visualisierung.[3] Der Einsatz computergestützter Verfahren und die Erweiterung des Methodenkatalogs eröffnen neue geisteswissenschaftliche Forschungsmöglichkeiten.[4] Digitale Ressourcen erweitern die Analyse von Forschungsdaten um quantitative Methoden und schaffen neue Einsichten, die vorher mühsam händisch erarbeitet werden mussten und damit nahezu unmöglich erschienen.[5] Diese forschungsorientierte Computerunterstützung ist der Grundgedanke der Digital Humanities.[6]

Nicht nur im akademischen Kontext, sondern auch in den Medien wird ausgiebig die Entstehung und Entwicklung der Digital Humanities diskutiert, deren interdisziplinäres Praxisfeld sich vor allem im letzten Jahrzehnt rasant entwickelt hat.[7] Voraussetzung für diese Entwicklung sind sowohl die fortschreitenden algorithmischen Möglichkeiten, die durch zunehmend leistungsfähigere und günstigere Hardware bedingt werden, als auch die Bestandserfassung, Informationsaufbereitung, -standardisierung und -bereitstellung sowie die Langzeitarchivierung durch Bibliotheken, Archive und Museen. Die Nutzung von digitalisierten Dokumenten und Objekten schafft eine enge Verbindung der Digital Humanities mit den Bibliotheken.[8] Wissenschaftliche Bibliotheken sind hierbei jedoch nicht bloße »Lückenfüller«[9] zwischen Geisteswissenschaften und Informatik, sondern vielmehr gleichberechtigte Kooperationspartner.[10] Als wichtige Säule dieser Zusammenarbeit zwischen Geisteswissenschaften und wissenschaftlichen Bibliotheken fungieren in diesem Kontext u. a. Fachreferentinnen und Fachreferenten, die angesichts grundsätzlicher Herausforderungen in Projekten der Digital Humanities unterstützen und »in großer Nähe zum fachlichen Kontext, zur Fachkultur und fachlichen Community«[11] stehen.

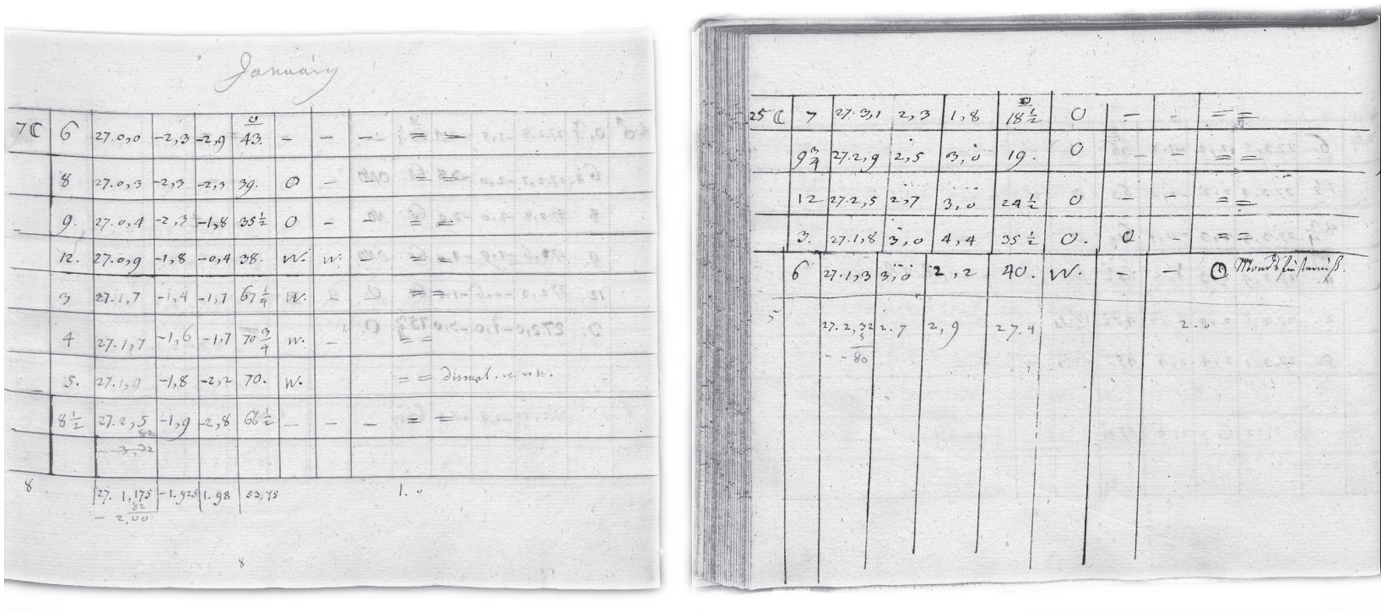
»Observationes meteorologicae« als meteorologische Forschungsdaten

Die UB Regensburg leitet derzeit ein Projekt zur computergestützten Erschließung und Aufbereitung wertvoller Schriftquellen. Seit der Übernahme der Sammlungen der Bibliothek der Philosophisch–Theologischen Hochschule Regensburg ist die Universitätsbibliothek im Besitz eines unikalen Bestands meteorologischer Aufzeichnungen aus dem 18. und 19. Jahrhundert.[12] Die handschriftlich geführten »Observationes meteorologicae« umfassen 53 Jahre und zählen zu den ältesten durchgängig geführten Wetteraufzeichnungen Europas.[13] Sie entstanden ab dem 1. Januar 1771 im Regensburger Kloster St. Emmeram und wurden dort bis zum 31. Dezember 1827 fortgeführt.[14] Das Kloster war zum damaligen Zeitpunkt sowohl aufgrund des dort tätigen Fachpersonals als auch aufgrund der Tatsache, dass das Kloster eine Bücherei, eine Wetterstation sowie eine Sternwarte beherbergte, ein fortschrittliches Zentrum naturwissenschaftlicher Forschung und, wie die Daten zeigen, in ein weites Netz monastischer Gelehrtheit eingebunden.

Coelestin Steiglehner, Professor für Mathematik und Naturwissenschaften, begann mit den Aufzeichnungen der Wetterdaten im Jahr 1771 und somit bereits neun Jahre vor der Gründung der meteorologischen Gesellschaft in Mannheim, welche anfangs mit Aufzeichnungen, Auswertungen und wissenschaftlichen Abhandlungen aus St. Emmeram versorgt wurde. Steiglehner, der die Aufzeichnungen der Wetterdaten für die folgenden sieben Jahre fortführte, hielt ab 1784 als erster deutscher Professor Vorlesungen über Wetterkunde an einer deutschen Universität, weshalb er auch den Titel »Vater der Meteorologie« trägt. Ab 1778 übernahmen Placidus Heinrich und anschließend seine Schüler die Anfertigung der Aufzeichnungen. Heinrich war ebenfalls ein anerkannter Naturwissenschaftler seiner Zeit und zwischen 1791 und 1798 Professor für Naturlehre, Stern- und Witterungskunde in Ingolstadt. Unter seine Leitung fiel in Regensburg der Einsatz standardisierter Instrumente der Akademie der Wissenschaften. Auf diese Weise konnten die Beobachtungen aus St. Emmeram in ein regionales Messnetzwerk eingebunden werden. Bis zu dreizehnmal am Tag wurden Temperatur, Luftdruck, Luftfeuchtigkeit, Windstärke und Windrichtung festgehalten, ab 1782 kamen weitere Informationen zur aktuellen Bewölkung, Nebel, Regen, Überschwemmungen, Mondphasen und Sonnenflecken hinzu. Mit den »Münchener Ephemeriden« veröffentlichte Heinrich 1797 die älteste Darstellung meteorologischer Daten Deutschlands und bewies darin präzises, wissenschaftliches Arbeiten. Nach seinem Tod 1825 führte der Naturwissenschaftler Ferdinand von Schmöger als Direktor der Emmeramer Sternwarte die Wettertagebücher bis zum 31. Dezember 1827 fort.[15]

Dank des hohen Maßes an Homogenität und Kontinuität eignen sich die »Observationes meteorologicae« damit in besonderer Weise für systematische, computergestützte Analyseansätze im Sinne der Digital Humanities. Mithilfe der erschlossenen Daten ist es möglich, meteorologische Momentaufnahmen zu rekonstruieren, um gesellschaftliche Auswirkungen und Folgen von Klimaänderungen auszumachen. Regional- und wirtschaftshistorische Archivbestände können zudem durch die Wetterdaten komplementiert werden. Das Zusammenwirken dieser unterschiedlichen Bestände wiederum kann zur Analyse lokalhistorischer Entwicklungen und Ereignisse herangezogen werden.[16]

Zur Erschließung von digitalisierten Schriftquellen werden mithilfe der *Optical Character Recognition* (OCR) optische Muster in digitalen Bildern zu maschinenlesbaren Texten transformiert. Zu den einzelnen Schritten der OCR zählen die Bildoptimierung des Digitalisats, die Analyse des Layouts sowie dessen Segmentierung in Textzeilen, die automatische Zeichenerkennung der segmentierten Textzeilen sowie die Kontrolle



1 »Observationes meteorologicae« – Wetteraufzeichnungen aus dem Jahr 1793
Foto: UB Regensburg

und Nachbereitung des erkannten Textes.[17] Während die automatische Erkennung moderner Drucke gegenwärtig nahezu fehlerfrei durchgeführt werden kann, stellen historische und besonders handschriftlich verfasste Dokumente weiterhin eine Herausforderung dar. Zum einen weisen historische Dokumente häufig beschädigte und somit fragmentierte Zeichen und Artefakte von dahinterliegenden Seiten auf. Zum anderen erschweren eine inkonsistente Seitengestaltung, grafische Abbildungen, kunstvolle Verzierungen sowie Ornamente, geschwungene Initialen und nicht zuletzt die Individualität der menschlichen Schrift die Erkennung.[18] Darüber hinaus mangelt es an linguistischen Tools zur Verbesserung der Qualität der Texterkennung, was auf fehlende orthografische Standards sowie kontinuierliche Änderungen der Rechtschreibung als auch auf die Ambiguität von Abkürzungen und Symbolen zurückzuführen ist. Eine zeit- und kostenaufwendige manuelle Erschließung der stetig wachsenden Menge an digitalisierten Archivalien, meist durch mindestens zwei Personen im sogenannten *Double-Keying*-Verfahren, ist bisher unausweichlich und aufgrund begrenzter Kapazitäten kaum zu bewältigen.[19]

Layout- und Texterkennung mithilfe neuronaler Netze

Trotz der vielseitigen Herausforderungen können auf dem Gebiet der Layout- und Texterkennung von historischen Dokumenten in den letzten Jahren große Fortschritte dank des Einsatzes neuronaler Netzwerke verzeichnet werden. Eine nahezu beliebige Anzahl von Rechenoperationen lässt sich in einem neuronalen Netzwerk in Form von Schichten aneinanderreihen, um Daten zu strukturieren und zu kategorisieren. Eine höhere Anzahl solcher Schichten führt häufig zu genaueren Ergebnissen und wird auch als *Deep Learning* bezeichnet. Manuell erstellte Beispieldaten, sogenannte Trainingsdaten, bringen dem Netzwerk bei, wie es bei der Strukturierung unbekannter Daten vorzugehen hat. Die erste praktische Anwendung von neuronalen Netzen stellte das automatisierte Auslesen von handgeschriebenen Postleitzahlen durch den *United States Postal Service* in den 1990er-Jahren dar.[20] Zunehmend leistungsfähigere und kostengünstigere Hardware sowie die wachsende Menge an verfügbaren Daten verhalfen neuronalen Netzwerken zum Einzug in sämtliche Anwendungsbereiche der Informatik. Durch den Einsatz neuronaler Netze konnte schließlich auch das seit den Anfängen der Handschriftenerkennung (HTR) bekannte Paradoxon von Kenneth Sayre[21] gelöst werden. Mehr als 30 Jahre lang wurden für die Texterkennung die Buchstaben der Wörter isoliert und separat mit Buchstabenvorlagen abgeglichen. Handschriftlicher Text mit zusammenhängenden Buchstaben erschwert diese Segmentierung, welche gemäß Sayres Paradoxon erst erfolgen kann, wenn zuvor das ganze Wort erkannt wurde. Dank des Einsatzes neuronaler Netze sind gängige Verfahren der Texterkennung nun nicht mehr zeichen-, sondern zeilenorientiert, weshalb eine Segmentierung der Buchstaben überflüssig und infolgedessen die Erkennung weniger anfällig für Zeichenvarianz ist.[22] Ebenfalls können mithilfe neuronaler Netze heterogene Seitenstrukturierungen historischer Dokumente mit einem generischen An-

satz erschlossen werden. Ferner kann die Text- und Layouterkennung mit individuellen Trainings an die unterschiedlichsten Arten von Dokumenten und Schriften angepasst werden.

Nach wie vor erfolgen sowohl die Erstellung erforderlicher Trainingsdaten als auch die Korrektur der Ergebnisse der OCR nicht automatisiert und müssen aus diesem Grund manuell vom Nutzer vorgenommen werden. Dennoch kann die Computerunterstützung den Transkriptionsprozess beschleunigen, wenn sie nutzerorientiert eingebettet wird.[23] Existierende Tools zum digitalen Transkribieren, die den langwierigen Transkriptionsprozess erleichtern sollen, besitzen zwar einen hohen Funktionsumfang, weisen aktuell aber teils schwerwiegende Defizite in der Handhabung und Benutzerfreundlichkeit auf. So wurde in einer eigenen Studie die Gebrauchstauglichkeit eines Standard-Tools zur Anfertigung von Transkriptionen evaluiert. Dabei zeigte sich, dass eine unzureichende Nutzerführung, unverständliche Fehlermeldungen sowie fehlendes Nutzerfeedback zur Unzufriedenheit der Nutzer beitragen. Nahezu alle Probanden scheiterten aufgrund des hohen Einarbeitungs- und Lernaufwands an der Erstellung einer digitalen Transkription.[24] Doch betrifft dies nicht nur digitale Transkriptionstools, denn in Projekten auf dem Gebiet der Digital Humanities wird der Usability während der Software-Entwicklung nur selten Aufmerksamkeit geschenkt.[25] Viele Tools können sich daher bei der breiten Masse an Nutzern nicht durchsetzen und bleiben nur wenigen Experten vorbehalten.[26]

Bezüglich der oben beschriebenen handschriftlich geführten »Observationes meteorologicae« erschwert des Weiteren die tabellarische Struktur, in der die Messwerte festgehalten wurden, die Erschließung der Aufzeichnungen. Sowohl die Lokalisierung als auch die semantische Erschließung von Tabellen stellten bisher keine Schwerpunkte in der OCR–Forschung dar. Bei der strukturellen Erkennung von Tabellen zeigen generische maschinell lernende Tools unter dem Einsatz neuronaler Netzwerke somit eklatante Schwächen. Das größte Problem stellt in diesem Zusammenhang die hohe Dichte an Objekten innerhalb einer Tabelle dar, aufgrund derer wichtige Details übersehen werden können.[27]

Erschließung von Tabellen an der UB Regensburg

Ausgehend von der beschriebenen Problematik soll an der UB Regensburg ein digitales Werkzeug nutzerorientiert entwickelt werden, das speziell auf die Erschließung von Tabellen ausgelegt ist. Den ersten Schritt in einem OCR–Prozess stellt, wie oben bereits beschrieben, die Optimierung des Bildmaterials dar. Dabei werden starke Verzerrungen der gescannten Dokumente ausgeglichen, der Kontrast wird angepasst, um den Durchschlag von Rückseiten zu reduzieren, und Ringartefakte sowie Rauschen werden entfernt. Im nächsten Schritt erfolgt der Einsatz sogenannter *Fully Convolutional Neural Networks* für die Erkennung der Seitenstruktur, einschließlich der Lokalisierung von Tabellen und Randnotizen. Diese Art von neuronalen Netzwerken ist in der Lage, antrainierte Muster in Grafiken zu finden. Diese Muster können wiederum mit bestimmten Wahrscheinlichkeiten Seitenelementen zugeordnet werden. Nach welchen Seitenelementen der Algorithmus suchen soll, kann durch annotierte Trainingsdaten festgelegt werden, in denen Seitenelemente farblich markiert und bezeichnet werden.[28] Neben Abbildungen, Ornamenten und Textzeilen lassen sich damit auch die Tabellen der Wettertagebücher lokalisieren. Für die anschließende Erschließung der Struktur wird aus den Schnittpunkten von horizontalen und vertikalen Linien auf die Tabellenzellen geschlossen. Um eine Linie als Tabellenlinien zu klassifizieren, werden die Rotation, Länge sowie der durchschnittliche Farbwert der Pixel, der eine bestimmte Schwärze aufweisen muss, herangezogen.[29] In einer Stichprobe mit 20 Testdokumenten konnten alle vorkommenden Tabellen lokalisiert und zu 87 % korrekt erschlossen werden.[30] Im Zuge der Tabellenlokalisierung wurde das Layout von insgesamt 30 Seiten annotiert und als Trainingsdaten für die Layouterkennung bereitgestellt. Weitere Maßnahmen zur Verbesserung des Kontrasts und der Reduzierung des Rauschens können dabei helfen, die strukturelle Erschließung zu verbessern.

Sowohl die Zellen der Tabellen als auch erkannte Textzeilen von Überschriften und Randnotizen stellen die Eingabe für die Texterkennung dar. Das von der Universität Würzburg entwickelte Softwaremodul *Calamari*[31] implementiert aktuelle Techniken zur Texterkennung. Hierbei wird eine Kombination aus *Fully Convolutional Neural Networks* und bidirektionalen *Long Short Term Memory Networks* für die Erkennung eingesetzt.[32] So wird eine Textzeile gleichzeitig jeweils vor- und rückwärts in horizontaler Richtung pixelweise abgetastet und es wird für jeden Abschnitt eine Wahrscheinlichkeit berechnet, mit der die Pixel Teil eines bestimmten Zeichens sind. Dabei werden Informationen aus den vorherigen und noch folgenden Abschnitten in die Entscheidung miteinbezogen. Anschließend ist es möglich, den vom Algorithmus prognostizierten Text durch den Einsatz linguistischer Tools zu optimieren. Im Vergleich zu anderen freiverfügbaren OCR–Programmen liefert *Calamari* die höchsten Erkennungsraten bei niedrigsten Laufzeiten und der geringsten Menge an notwendigen Trainingsdaten.

Bisher wurde der Algorithmus mit 30 Seiten manuell transkribierter Wetterdaten trainiert, was 1.246 Tabellenzellen inklusive Überschriften und Randnotizen entspricht.[33] Die Fehlerrate liegt aktuell noch bei 72,55 %, was die Erstellung weiterer Trainingsdaten notwendig macht. Eine Schätzung bezüglich der Anzahl erforderlicher Trainingsdaten kann aufgrund des hohen Individualitätsgrades handschriftlicher Bestände nicht pau-

schal angegeben werden. Beispielsweise rät jedoch die Benutzeranleitung der Transkriptionssoftware *Transkribus* dazu, 5.000–15.000 Wörter für das Training zu verwenden.[34]

Ausblick

Das digitale Werkzeug, das an der UB Regensburg nutzerorientiert entwickelt wird und das speziell auf die Erschließung von Tabellen ausgelegt ist, soll am Ende des Projekts Teil einer plattformunabhängigen Desktop- und Webanwendung werden. Neben den technischen Anforderungen sollen auch die Nutzerbedürfnisse spezifiziert und die Usability im Entwicklungsprozess berücksichtigt werden. So wurde im ersten Schritt bereits der Problemkontext analysiert, und es wurden Annahmen über Nutzer formuliert. Mithilfe sogenannter *Personas* lassen sich die Motivationen, Bedürfnisse, Ziele und die Verhaltensweisen potenzieller Nutzer idealtypisch herleiten.[35] Da es sich zunächst um bloße Annahmen handelt, werden in darauffolgenden Interviews mit Editoren, geisteswissenschaftlichen Studierenden und Bibliotheksmitarbeitenden diese Annahmen evaluiert und nachgebessert. Um die Usability bestehender Tools zu analysieren, kann unter anderem eine heuristische Evaluation durchgeführt werden, im Rahmen derer Verstöße gegen etablierte Design- und Usability-Standards bei der Nutzung des zu testenden Tools dokumentiert werden.[36] Darüber hinaus gewähren Usability-Tests, in denen potentielle Nutzer bei der Bearbeitung typischer Aufgaben beobachtet werden, Einblicke in das konkrete Nutzerverhalten.[37]

Die gewonnenen Erkenntnisse dienen im Anschluss als Grundlage für den Entwurf von ersten interaktiven Prototypen. Diese werden in regelmäßigen Abständen in Zusammenarbeit mit potenziellen Nutzern evaluiert, um Verbesserungsvorschläge frühzeitig einarbeiten zu können.[38] Derzeit werden im Rahmen des an der UB Regensburg laufenden Projekts erste Prototypen des Programms auf Grundlage von Nutzerfeedback evaluiert. Zudem wird daran gearbeitet, in Gesprächen mit geisteswissenschaftlichen Forschenden Fachbegriffe und Methoden des Transkriptionsprozesses zu vertiefen. In einem Projektseminar soll im Sommersemester 2020 in Zusammenarbeit mit der Geschichtswissenschaft der Universität Regensburg die Anwendung in einem praktischen Kontext evaluiert werden. Die Ergebnisse werden im zweiten Teil dieses Beitrags präsentiert und diskutiert. Um Transparenz und Nachhaltigkeit zu garantieren, soll der Quellcode der Anwendung nach Ende der Projektlaufzeit unter der *GNU General Public License* frei zugänglich sein.

Ferner ist es Ziel, die Wetterdaten unter Einsatz der Software konsistent zu erschließen und in dem etablierten Standardformat TEI zu erfassen. In diesem Zusammenhang bietet sich für die Repräsentation digitaler Transkriptionen das PAGE-Format an. Eine Auswahl an vielfältigen Seitenelementen sowie die Dokumentation von angewandten Bildoptimierungsschritten ermöglichen eine detaillierte und nachvollziehbare Darstellung des Transkriptionsprozesses.[39] Die PAGE-Dateien der jeweiligen Dokumentseiten können in das METS–Schema eingebettet werden, das als Containerformat Metadaten zu einer kompletten Sammlung umfasst. Die Nachnutzung und Verfügbarkeit der Daten für verschiedenste Nutzungsszenarien soll dadurch garantiert und die Bearbeitung eventuell noch nicht absehbarer Forschungsfragen unterstützt werden. Der Publikationsserver der UB Regensburg[40] eignet sich dabei als Plattform, um Forschenden Zugang zu den Daten zu ermöglichen.

Abschließend ist die interdisziplinäre Relevanz der ursprünglich handschriftlich erfassten »Observationes meteorologicae« nochmals hervorzuheben. Diese Relevanz wird dank der Aufbereitung der Daten in maschinenlesbarer Form und der damit möglich gemachten Analysen mithilfe digitaler Methoden weiter zunehmen. Dennoch wird im Rahmen des Projekts zur Erschließung der »Observationes meteorologicae« deutlich, dass die computergestützte Layout- und Texterkennung handschriftlicher Aufzeichnungen nach wie vor nicht vollautomatisiert und ohne menschliche Interaktionen durchgeführt werden kann. Dies ist hauptsächlich der Menge an manuell zu erstellenden Trainingsdaten geschuldet, die für eine fehlerfreie Texterkennung notwendig ist. Wie in dem vorliegenden Beitrag dargestellt, wird aktuell jedoch viel Forschung auf dem Gebiet der Handschriftenerkennung betrieben und damit sowohl der Trainings- als auch der Erkennungsprozess stetig optimiert. In Verbindung mit einem nutzerzentrierten Entwicklungsprozess und Tools, die weniger auf Experten, sondern mehr auf Erst- und Gelegenheitsnutzer ausgelegt sind, eröffnen sich in Zukunft vielversprechende Möglichkeiten, um die Masse an digitalisierten Schriftquellen effizient erschließen und so die bestandsbezogene Forschung voranbringen zu können.

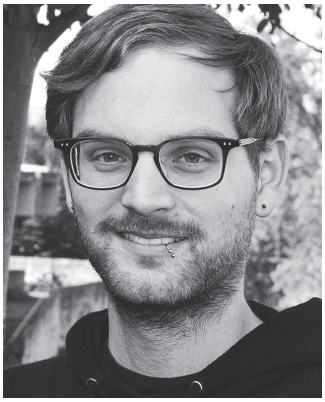
Anmerkungen

1 Vgl. Heike Neuroth: Bibliothek, Archiv, Museum. In: Fotis Jannidis, Hubertus Kohle & Malte Rehbein (Hrsg.). *Digital Humanities: Eine Einführung*. Stuttgart: J. B. Metzler Verlag 2017, S. 213–222, hier S. 213 f.

2 Vgl. Thomas Stäcker: Die Sammlung ist tot, es lebe die Sammlung! – Die Digitale Sammlung als Paradigma moderner Bibliotheksarbeit. In*: Bibliothek Forschung und Praxis* 43/2 (2019), S. 304–310, hier S. 307.

3 Vgl. Anm. 2, hier S. 304.
4 Vgl. Gerhardt Lauer: Die digitale Vermessung der Kultur: Geisteswissenschaften als Digital Humanities. In: Heinrich Geiselberger & Tobias Moorstedt (Hrsg.). *Big Data: Das neue Versprechen der Allwissenheit.* Suhrkamp (2013), S. 99–116, hier S. 110.
5 Vgl. Lorna Hughes, Panos Constantopoulos & Costis Dallas: Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In: Susan Schreibman, Ray Siemens & John Unsworth (Hrsg.). *A New Companion to Digital Humanities*. Wiley Blackwell 2016, S. 150–170, hier S. 153.
6 Vgl. Anm.4, hier S. 101.
7 Vgl. Guido Koller: *Geschichte digital: Historische Welten neu vermessen*. W. Kohlhammer GmbH 2016, hier S. 21. Für eine ausführliche Beschreibung der Geschichte und des aktuellen Stands der Digital Humanities, vgl. Anm. 10.
8 Vgl. Wolfram Horstmann: Zur Rolle von Bibliotheken in digitalen Forschungsinfrastrukturen. In: Achim Bonte & Juliane Rehnolt (Hrsg.). *Kooperative Informationsinfrastrukturen als Chance und Herausforderung* (2018), S. 93–109, hier S. 99.
9 Vgl. Pascal-Nicolas Becker & Fabian Fürste: *Sollen wir Bibliothekare jetzt alle Informatiker werden?* Verfügbar unter: https://b-u-b.de/sollen-wir-bibliothekare-jetzt-alle-informatiker-werden/
10 Vgl. Frédéric Döhl: Digital Humanities und Bibliotheken: Über technisch-organisatorische Infrastruktur hinausgedacht. In: *ZfBB* 66/1 (2019), S. 4–18, hier S. 5 f.
11 Vgl. Inga Tappenbeck: Wissenschaftlicher Dienst im Wandel? Eine Bestandsaufnahme am Beispiel der Universitätsbibliotheken in Nordrhein-Westfalen. In: Haike Meinhardt & Inga Tappenbeck (Hrsg.). *Die Bibliothek im Spannungsfeld: Geschichte – Dienstleistungen – Werte* (2019), S. 129–140, hier S. 132.
12 Vgl. Andreas Becker: Die Schriftgutverwaltung des Lyzeums Albertinum und der Philosophisch-Theologischen Hochschule im Spiegel der Überlieferung im Universitätsarchiv Regensburg. In: *Verhandlungen des Historischen Vereins Regensburg* Bd. 154 (2014), S. 275–292, hier S. 285.
13 Universitätsbibliothek Regensburg: *Observationes meteorologicae: Placidus Heinrich und seine Wetteraufzeichnungen* (2010). Verfügbar unter: http://bibliothek.uni-regensburg.de/meteorologie/
14 Die ersten drei Jahre der Aufzeichnungen sind nicht im Bestand der Universitätsbibliothek enthalten.
15 Vgl. Martina Lorenz: Naturforschung in St. Emmeram. In: *Im Turm, im Kabinett, im Labor. Streifzüge durch die Regensburger Wissenschaftsgeschichte*. Universitätsverlag Regensburg 1995, S. 12–29.
16 Vgl. Constantin Lehenmeier & Manuel Burghardt: Historische Wetterdaten im Spannungsfeld zwischen OCR und User-Centered Design. In Manuel Burghardt & Claudia Müller-Birn (Hrsg.). *INF-DH-2018*. Bonn: Gesellschaft für Informatik e.V. 2018, hier S. 1 f.
17 Vgl. Malte Rehbein: Digitalisierung. In: Fotis Jannidis, Hubertus Kohle & Malte Rehbein (Hrsg.). *Digital Humanities: Eine Einführung*. Stuttgart: J. B. Metzler Verlag 2017, S. 179–198, hier S. 194 f.
18 Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, Frank Puppe: *OCR4all – An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings* (2019).
19 Vgl. Anm. 17, hier S. 193.
20 Vgl. Francois Chollet: *Deep Learning with Python*. Manning Publications 2017, hier S. 15.
21 Vgl. Kenneth M. Sayre: Machine recognition of handwritten words: A project report. In: *Pattern Recognition* 5/3 (1973), S. 213–228.
22 Vgl. Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke & Jürgen Schmidhuber: A Novel Connectionist System for Unconstrained Handwriting Recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31/5 (2008), S. 855–868.
23 Vgl. LWC van Lit: *Among Digitized Manuscripts. Philology, Codicology, Paleography in a Digital World*. Brill 2019, hier S. 117.
24 Vgl. Constantin Lehenmeier & Manuel Burghardt: Usability statt Frustration. In: Claude Draude, Martin Lange & Bernhard Sick (Hrsg.). *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft (Workshop-Beiträge)*. Bonn: Gesellschaft für Informatik e.V. 2019, S. 97–106.
25 Vgl. Klaus Thoden, Juliane Stiller, Natasa Bulatovic, Hanna-Lena Meiners & Nadia Boukhelifa: User-Centered Design Practices in Digital Humanities-Experiences from DARIAH and CENDARI. In: *ABI Technik* 37/2 (2017). Verfügbar unter: https://www.degruyter.com/view/j/abitech.2017.37.issue-1/abitech-2017-0002/abitech-2017-0002.xml
26 Vgl. Liza Potts: Archive Experiences: A Vision for User-Centered Design in the Digital Humanities. In: Jim Ridolfo & William Hart-Davidson (Hrsg.). *Rhetoric and the Digital Humanities*. University of Chicago Press 2015, S. 255–263, hier S. 255 f.
27 Vgl. Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel & Sheraz Ahmed: DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (2017).
28 Vgl. Sofia Ares Oliveira, Benoit Seguin & Frederic Kaplan: dhSegment: A generic deep-learning approach for document segmentation. In: *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (2018).
29 Vgl. Benjamin Charles Germain Lee: Line detection in binary document scans: A case study with the international tracing service archives. In: *IEEE International Conference on Big Data (Big Data)* (2017).
30 Für die Evaluierung wurden Tabellen manuell transkribiert und den auszuwertenden Ergebnissen gegenübergestellt. Hierzu werden die jeweiligen Positionen der Tabellen mithilfe des Jaccard-Koeffizienten verglichen. Um die Struktur, also die Anordnung der Zellen, miteinander vergleichen zu können, werden die Zellen und deren Spalten- und Zeilennummerierung in HTML umgewandelt und die Übereinstimmung durch das BLEU-Maß ausgewertet.
31 Vgl. https://github.com/Calamari-OCR/calamari
32 Vgl. Christoph Wick, Christian Reul & Frank Puppe: *Calamari – A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition* (2018).
33 Die Trainingsdaten wurden mithilfe der Software *Transkribus* erstellt.
34 Vgl. https://transkribus.eu/wiki/images/2/21/Modell_Training_in_Transkribus.pdf
35 Vgl. Leah Buley: *The User Experience Team of One. A Research and Design Survival Guide*. Rosenfeld Media 2013, hier S. 132–135.
36 Vgl. Anm. 35, hier S. 136–139. Als Heuristiken wurden die Grundsätze des Interaktionsdesigns von Bruce Tognazzini ausgewählt. Verfügbar unter: https://asktog.com/atc/principles-of-interaction-design/
37 Vgl. Jeff Sauro & James R. Lewis: *Quantifying the User Experience: Practical Statistics for User Research*. 2. Aufl. Morgan Kaufmann 2016, hier S. 10.

38 Vgl. Anm. 16, hier S. 3.
39 Vgl. Stefan Pletschacher & Apostolos Antonacopoulos: The PAGE (Page Analysis and Ground-truth Elements) Format Framework. In: *International Conference on Pattern Recognition* (2010), S. 257–260, hier S. 257.
40 Vgl. https://epub.uni-regensburg.de/



Der Verfasser
Constantin Lehenmeier, Universitätsbibliothek Regensburg, Universitätsstraße 31, 93053 Regensburg, constantin.lehenmeier@ur.de
Foto: privat