ARCHEION
2021, T. CXXII, s. 129–153
ISSN 0066-6041
e-ISSN 2658-1264

DOI: 10.4467/26581264ARC.21.006.14486

STUDIA I MATERIAŁY

Transkribus im Archiv – Ein polnisch-deutsches Projekt zur Handschriftentexterkennung an historischen Dokumenten

Dirk Alvermann

Universitätsarchiv Greifswald / University Archives Greifswald (Germany)
alverman@uni-greifswald.de

Paweł Gut

Uniwersytet Szczeciński / University of Szczecin (Poland)
pawel_gut@wp.pl, ORCID 0000-0003-3148-3298

Transkribus in the archives – a Polish-German project of reading historical documents

ABSTRACT
Even 10 years ago, the idea that historical manuscripts, regardless of time of creation and origin, could be „read" and searched using automated processes seemed unrealistic. However, thanks to modern machine learning methods and the use of artificial intelligence, it is now possible. Following the development of Transkribus platform (http://transkribus.eu/), a tool has become available that allows free open access to this technology. Handwriting recognition permits automatic conversion of large numbers of historical manuscripts into fully legible texts. This development will influence and change the work of archives over the next several years, especially with regard to how their collections are made accessible digitally. Using the example of a Polish-German cooperation project, the article presents the use of handwriting recognition technology in the context of an archival digitisation project and discusses the technical requirements, technological work input and results of using Transkribus in an archive.

KEYWORDS
handwriting recognition, Transkribus, documents, digitalization, Pomerania, Polish-German cooperation

Transkribus w archiwum – polsko-niemiecki projekt odczytania dokumentów historycznych

STRESZCZENIE
Jeszcze 10 lat temu pomysł, że rękopisy historyczne, niezależnie od czasu i pochodzenia, można „czytać" i przeszukiwać za pomocą zautomatyzowanych procesów, wydawał się nierealny. Dzięki nowoczesnym metodom uczenia się maszynowego i wykorzystaniu sztucznej inteligencji jest to obecnie możliwe. Wraz z rozwojem platformy Transkribus (http://transkribus.eu/) dostępne jest narzędzie, które pozwala na otwarty dostęp do tej technologii. Rozpoznawanie pisma ręcznego umożliwia automatyczną konwersję dużej liczby rękopisów historycznych na w pełni czytelne teksty. Ten rozwój wpłynie i zmieni pracę archiwów w perspektywie kilkunastu

SŁOWA KLUCZOWE
rozpoznawanie pisma ręcznego, Transkribus, dokumenty, digitalizacja, Pomorze, współpraca polsko-niemiecka

lat, zwłaszcza sposób cyfrowego udostępniania ich zbiorów. Na przykładzie polsko-niemieckiego projektu współpracy, w artykule przedstawiono wykorzystanie technologii rozpoznawania pisma ręcznego w kontekście projektu digitalizacji archiwalnej oraz omówiono wymagania techniczne, wkład prac technologicznych i rezultaty wykorzystania Transkribusa w archiwum.

Das Projekt

Im Oktober 2020 begann ein von der Historischen Kommission für Pommern angeregtes und auf sieben Jahre angelegtes Projekt unter dem Titel „Klosterregister und Klosterbuch für Pommern". Es ist an der Christian-Albrechts-Universität zu Kiel innerhalb der Forschungsstelle *Geschichte und kulturelles Erbe der Klöster und Stifte im Ostseeraum bis zur Reformation* angesiedelt und wird von dort aus koordiniert und geleitet[1]. In diesem Forschungsprojekt wird die Geschichte und das Wirken aller 64 Klöster, Stifte, Konvente und Kommenden in Vor- und Hinterpommern historisch erfasst. Der Untersuchungsraum erstreckt sich dabei grenzüberschreitend auf das Bundesland Mecklenburg-Vorpommern und die Wojewodschaften Westpommern/województwo zachodniopomorskie und Pommern/województwo pomorskie in der Republik Polen. Die Arbeiten erfolgen in enger Kooperation mit deutschen und polnischen sowie dänischen und schwedischen Forschungseinrichtungen. Der Untersuchungszeitraum reicht von den Anfängen der Klöster und Stifte im 12. Jahrhundert bis zu ihrer Aufhebung im Zuge der Reformation. In einem ersten Projektabschnitt, dem *Klosterregister für Pommern*, soll zu jeder Niederlassung ein möglichst auf Vollständigkeit hin angelegtes Literatur-, Quellen-, Denkmal- und Inventarverzeichnis erstellt werden. In einem zweiten Projektabschnitt werden die gewonnenen Ergebnisse den Autorinnen und Autoren für die Abfassung von wissenschaftlichen Artikeln zur Geschichte der einzelnen Klöster und Niederlassungen zur Verfügung gestellt. Zum Projektende soll daraus ein „Klosterbuch für Pommern" entstehen, das in der Tradition des bekannten Werkes von Hermann Hoogeweg *Stifter und Klöster der Provinz Pommern*[2] steht und die traditionelle Beschreibung der pommerschen

[1] CAU Christian-Albrechts-Universität zu Kiel, Das Klosterregister und Klosterbuch für Pommern, https://www.histsem.uni-kiel.de/de/das-institut-1/abteilungen/regionalgeschichte-mit-schwerpunkt-schleswig-holstein/projekte/pommersches-klosterbuch [abgerufen am 28.6.2021].

[2] Hermann Hoogeweg, *Die Stifter und Klöster der Provinz Pommern*, Bd. 1–2, Stettin 1924–1925.

„Klosterlandschaft" mit neuesten Erkenntnissen archivischer Quellenforschung, archäologischer Grabungen und kunstgeschichtlicher Untersuchungen kombiniert.

Als Teilprojekt des Klosterregisters und Klosterbuchs für Pommern werden in Kooperation mit dem Staatsarchiv Stettin, dem Universitätsarchiv Greifswald und der Universitätsbibliothek Greifswald 2020–2021 die Urkundenregister der pommerschen Kirchen und Klöster von Hermann Hoogeweg, die im Staatsarchiv Stettin archiviert sind, digitalisiert und in der Digitalen Bibliothek Mecklenburg-Vorpommerns präsentiert. Damit soll der späteren Forschung zur Geschichte der einzelnen Klöster ein wichtiges Rechercheinstrument in zeitgemäßer Form zugänglich gemacht werden. Während das Staatsarchiv Stettin die Digitalisierung dieser bedeutenden Überlieferung übernommen hat, wird im Universitätsarchiv Greifswald eine Volltexterkennung der 34 Regestenbände durchgeführt. Dabei kommt eine neue Technologie – Handwritten Text Recognition (HTR) – zum Einsatz, die es erlaubt, die handschriftlichen Quellen mit Unterstützung Künstlicher Intelligenz automatisch zu entziffern und in einen durchsuchbaren Text umzuwandeln. Anschließend werden die Digitalisate und Volltexte durch die Universitätsbibliothek Greifswald in der Digitalen Bibliothek Mecklenburg-Vorpommerns in durchsuchbarer Form präsentiert[3]. Im Folgenden sollen die Voraussetzungen, Inhalte und technischen Details dieses Projektes kurz umrissen werden. Dabei wird auch allgemein auf die Möglichkeiten eingegangen, die die moderne Handschriftentexterkennung, als eine künftige „Schlüsseltechnologie" für die Archive eröffnet.

Warum Handschriftentexterkennung?

Schaut man auf die Zielgruppe klassischer Digitalisierungsangebote von Archiven, dann stellt man fest, dass sie sich häufig an dieselben Adressaten richten, die auch unsere Lesesäle aufsuchen, d.h. an Menschen die nicht nur über besondere Interessen, sondern auch über besondere Fähigkeiten verfügen,

[3] Mittlerweile stehen dort bereits über 10 Regestenbände für die Recherche zur Verfügung: Die Digitale Bibliothek Mecklenburg-Vorpommern, Regesten zu den Urkunden der pommerschen Kirchen und Klöster - Regesty dokumentów kościołów i klasztorów pomorskich, https://www.digitale-bibliothek-mv.de/viewer/toc/PPNAPSzczecinie_65_78_0_3_1/ [abgerufen am 5.7.2021].

z. B. das Lesen alter Handschriften. Wir reden dabei also von einem winzigen Bruchteil der Bevölkerung. Für alle übrigen sind die Digitalisate historischer Akten oder Urkunden eigentlich nur schön anzusehen – ihr konkreter Inhalt bleibt ihnen weitgehend unzugänglich (sei es nun, weil sie die Schrift nur schwer entziffern können oder weil es sprachliche Barrieren gibt).

Hält man sich das vor Augen, versteht man, warum Handschriftentexterkennung (HTR) in der Geschichte der digitalen Nutzung aber auch in der Erschließung von Archivalien ein völlig neues Kapitel aufschlägt. Mit einem Satz könnte man sagen: HTR gestattet den Schritt von der einfachen Digitalisierung zur digitalen Transformation archivalischer Quellen. Dank der HTR wird nämlich nicht nur das digitale Abbild einer Handschrift sondern auch ihr Inhalt automatisiert in einer für jedermann lesbaren und von Maschinen durchsuchbaren (und auch übersetzbaren) Form – und zwar über hunderttausende Seiten hinweg – verfügbar gemacht.

Damit ist nicht nur der Kreis der Laienforscher oder die Citizen Science im Allgemeinen angesprochen. Auch für wissenschaftliche Fachvertreter aus Disziplinen, in denen historische Hilfswissenschaften nicht zum klassischen Ausbildungskanon gehören, wird die Zugänglichkeit zu den Inhalten der Quellen erleichtert. Neue Konstellationen interdisziplinären Forschens werden ermöglicht. Und schließlich: da die Inhalte der Handschriften nun maschinell auswertbar sind, lassen sich Fragestellungen und Methoden der Digital Humanities weitaus leichter an das Material herantragen als zuvor.

Im Hinblick auf die Digitalisierung ermöglicht HTR den Archiven einen Schritt zu gehen, den die Bibliotheken weltweit dank OCR (Optical Character Recognition) und der damit verbundenen Volltextdigitalisierung historischer Buchbestände schon vor einem Jahrzehnt unternommen haben und der die Funktion und Nutzung digitaler Angebote der Bibliotheken dramatisch verändert hat. Beispielhaft dafür kann das Projekt „Austrian Books Online“[4] stehen, das zwischen 2011 und 2019 den gesamten urheberrechtsfreien Buchbestand der Österreichischen Nationalbibliothek zwischen 1501 und der 2. Hälfte des 19. Jahrhunderts (das sind 600.000 Werke mit 200 Millionen Druckseiten) digitalisiert und dank OCR auch für eine Volltextsuche online verfügbar gemacht

[4] Österreichische Nationalbibliothek, Austrian Books Online, https://www.onb.ac.at/digitaler-lesesaal/austrian-books-online-abo [abgerufen am 24.6.2021].

hat. Das sind Perspektiven, die im Bereich des Archivwesens und der historischen Handschriften bislang als technisch unrealistisch galten.

Nachdem sich das Bewusstsein davon, was mit HTR möglich ist, zunächst im Bereich der universitären Forschung verbreitete und dort in verschiedenen Projekten erprobt wurde, haben sich in vielen europäischen Ländern mittlerweile aber auch die Archive des Themas angenommen und erkunden aus unterschiedlichen Perspektiven die Einsatzmöglichkeiten von HTR im klassischen Aufgabenspektrum ihrer Häuser.

Das National Archief der Niederlande wird noch 2021 ein Portal starten, in dem die ersten zwei Millionen Seiten retrodigitalisierter Akten dank HTR-generierter Volltexte durchsuchbar sein werden[5]. Das Stadtarchiv Amsterdam hat bereits über 300.000 Seiten Notariatsakten seiner Bestände online durchsuchbar gemacht[6]. Das Staatsarchiv Zürich ist im Begriff 1000 Bände der Züricher Ratsmanuale (1484–1798) ebenfalls mit HTR-generierten Volltexten verfügbar zu machen[7]. In Mecklenburg-Vorpommern stellen das Landesarchiv, das Stadtarchiv Wismar und das Universitätsarchiv Greifswald insgesamt 225.000 Seiten durch HTR volltexterschlossene Gerichtsakten online zur Verfügung[8]. Das Finnische Staatsarchiv bietet 800.000 Seiten volltextdigitalisierter Hofgerichtsakten über ein besonderes Internetportal zur Recherche an[9] und plant die Bereitstellung von 1,5 Millionen Seiten finnischer Steuerrollen des 19. Jahrhunderts auf demselben Weg[10]. In Lissabon wird ein Projekt zur Digitalisierung von etwa 7 Millionen Seiten von Protokollen der Inquisition aus dem Archivo Nacional, die durch HTR erschlossen werden sollen, vorbereitet. Die Liste könnte noch fortgesetzt werden.

---

5 Zoeken in transcripties, www.zoekintranscripties.nl und den Vortrag von L. Keyser auf YouTube, 02 Transkribus in practise – Transkribus User Conference 02/2020, https://www.youtube.com/watch?v=xQPcJHGn8cM&t=866s [abgerufen am 28.6.2021].

6 Transkribus, Amsterdam notarial deeds, https://transkribus.eu/r/notarial/ [abgerufen am 28.6.2021].

7 De Gruyter, Digitalisierungsprojekte des Staatsarchivs Zürich mit Einsatz von Machine-Learning-Verfahren, https://www.degruyter.com/document/doi/10.1515/abitech-2020-2018/html [abgerufen am 5.7.2021].

8 Rechtsprechung im Ostseeraum, https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/ [abgerufen am 5.7.2021].

9 Search Finnish Court Records. Search and browse district court records from 1810 to 1870, https://tuomiokirjat.narc.fi/en [abgerufen am 28.6.2021].

10 Semantic Computing Research Group (SeCo), Handwritten Text Recognition in the Archives https://seco.cs.aalto.fi/events/2018/2018-10-23-heldig-summit/presentations/06-kallio.pdf [abgerufen am 28.6.2021].

In den hier genannten Projekten wurde und wird HTR als Mittel zur Erschließung „sperriger" Aktenbestände eingesetzt, die andernfalls nur durch langjährige kleinteilige Verzeichnung in einer annähernd vergleichbaren Tiefe hätten erschlossen werden können. Bestände von ähnlichem Charakter, bspw. die neuzeitlichen Protokollserien städtischer Magistrate oder anderer Behörden, bieten sich für diese Art der alternativen Erschließung förmlich an[11].

Das HTR darüber hinaus Potentiale in der Erschließungsarbeit hat, zeigen Beispiele wie die Digitalisierung und Erschließung der sogenannten „Vreemdelingskaarten" der Polizeidirektion im Amsterdamer Stadtarchiv[12]. Darin wurden während des zweiten Weltkriegs über 200.000 Flüchtlinge und Migranten erfasst, deren Namen und Daten nun Dank HTR automatisch erschlossen werden konnten.

Auch im Bereich der Retrokonversion von Findmitteln kann HTR neue Akzente setzen. So wurden 2019 im Staatsarchiv Zürich analoge Findmittel in Karteiform, darunter 30.000 Urkundenregesten zunächst für interne Zwecke durch HTR erschlossen, die später in das Archivinformationssystem übertragen werden sollen[13]. Ein ähnliches Projekt verfolgen das Staatsarchiv Stettin und seine Partner. Dabei sollen die Urkundenregister der pommerschen Kirchen und Klöster von Hermann Hoogeweg mit Hilfe der HTR in moderne Schrift „übersetzt" und anschließend in Verbindung mit den Digitalisaten der Originalhandschriften im Internet zur Recherche zur Verfügung gestellt werden.

Das Material – die Urkundenregister der pommerschen Kirchen und Klöster von Hermann Hoogeweg

Der Bestand an Pergament- und Papierurkunden des preußischen Staatsarchivs Stettin wurde seit der Entstehung dieser Einrichtung sukzessiv durch weitere Zugänge erweitert und erreichte zu Beginn des 20. Jh. den Umfang von 7000–8000 Schriftstücken. Die meisten von ihnen waren Urkunden, die aus

11 Beispielhaft wäre hier die Erschließung der Züricher Ratsprotokolle von 1484–1798 oder auch die Schweizer Bundesratsprotokolle zu nennen. An der Universität Tübingen ist ein Modellprojekt anhand der akademischen Senatsprotokolle durchgeführt worden und im Stadtarchiv Bautzen wird die komplette Serie der historischen Ratsprotokolle künftig online durchsuchbar sein.

12 Gemeente Amsterdam, Vreemdelingenkaarten, https://www.amsterdam.nl/stadsarchief/nieuws/vreemdelingenkaarten/ [abgerufen am 28.06.2021].

13 De Gruyter, Digitalisierungsprojekte des Staatsarchivs Zürich, op. cit.

ehemaligen geistlichen Archiven stammten: von Bischöfen, Kapiteln, Klöstern und Kirchen. Mit diesen Unterlagen haben sich im 19. Jh. Archivare wie Friedrich Ludwig von Medem, Robert Klempin, Gustav Kratz, Rodgero Prümers und Max Bär befasst. Teilweise bestanden diese Arbeiten in Recherchen zum *Pommerschen Urkundenbuch* (PUB), das seit den 1860er Jahren von den Mitarbeitern des Staatsarchivs Stettin erarbeitet worden ist[14].

Zu den von den Archivaren im 19. Jh. erstellten Findhilfsmitteln gehörten Inventare mit Kurzregesten. Sie hatten die Form von Büchern bzw. Karteien. Einige der Dokumente waren nicht inventarisiert, sondern verfügten nur über Kurzbeschreibungen auf den Rückseiten von Pergamentblättern. Darüber hinaus wurden die im Stettiner Archiv aufbewahrten Urkunden durch Beschreibungen und Regesten in den nachfolgenden Bänden der PUB (Bände 1–6), die pommersche Urkunden bis 1325 enthalten, archivalisch nachgewiesen.

Ein Wandel in der Bearbeitung mittelalterlicher Urkunden, insbesondere kirchlicher Provenienz, vollzog sich, als Hermann Moritz Rudolf Robert Hoogeweg (1857–1930) Direktor des Staatsarchivs in Stettin wurde. Dieser erfahrene Archivar spezialisierte sich auf die Geschichte kirchlicher Institutionen sowie auf die Erarbeitung von Findmitteln zu Schriftstücken aus der archivischen Überlieferung dieser geistlichen Anstalten. Er gab u.a. die Urkunden des Bistums Minden sowie ein Verzeichnis der Stifter und Klöster Niedersachsens heraus[15].

Nach seinem Amtsantritt am 1. April 1913 inspizierte Hermann Hoogeweg das gesamte Archivgut. Daraus entstand ein *Repertorium-Verzeichnis* für die im Stettiner Archiv aufbewahrten Bestände und Sammlungen[16]. Diese Arbeit war die Grundlage für die Feststellung, dass die Bestände/Sammlungen von klösterlichen und anderen kirchlichen Urkunden nur über unzureichende Archivrepertorien verfügten, die weder den Bedürfnissen der Benutzer des Archivs noch der Forscher zur mittelalterlichen Geschichte gerecht werden konnten. Diese Beobachtung fand die Aufmerksamkeit des Generaldirektors

---

[14] Erstes Band ist im 1868 publizirt. Herausgeber war Robert Klempin. *Pommersches Urkundenbuch*, Bd. 1 Abt. 1, 786–1253. Bearbeitet und herausgegeben von R. Klempin, Stettin 1868.

[15] M. Szukała, *Archiwum Państwowe w Szczecinie w latach 1914–1945. Ludzie i działalność*, Szczecin 2019, S. 26–29; H. Hoogeweg, *Verzeichnis der Stifter und Klöster Niedersachsens vor der Reformation, umfassend die Provinz Hannover, die Herzogtümer Braunschweig und Oldenburg, die Fürstentümer Lippe-Detmold und Schaumburg-Lippe, die Freien Städte Bremen und Hamburg und Hessisch-Schaumburg*, Hannover 1908.

[16] Archiwum Państwowe w Szczecinie (weiter AP Szczecin), Bestand: 65/78/0 Archiwum Państwowe w Szczecinie (Staatsarchiv Stettin) [1500] 1831–1945 [1971], Sig. 65/78/0/2/1116.

der Preußischen Staatsarchive, Professor Reinhold Koser, der anordnete, den gegebenen Zustand der Verzeichnung der ältesten Bestände im Stettiner Archiv zu verbessern[17]. Bereits bei Arbeiten an den Regesten stellte Hoogeweg fest, dass einige Urkunden aus dem 12. bzw. 13. Jh. aufgrund von unzulänglichen Archivinventaren nicht in den bereits erschienenen Bänden des Pommerschen Urkundenbuches veröffentlicht worden waren[18].

Hermann Hoogeweg, der im Rahmen seiner archivarischen Tätigkeit Erfahrungen mit der Erschließung von mittelalterlichem Archivgut hatte, wurde vom Generaldirektor R. Koser beauftragt, neue Findhilfsmittel zu den Urkunden kirchlicher Anstalten vom Zeitpunkt ihrer Entstehung im Pommern im 12. Jh. bis zur Einführung der Reformation im 16. Jh. zu erarbeiten. Diese Arbeiten wurden in den Jahren 1913–1923 durchgeführt. Im Endergebnis entstanden 34 Bücher mit Urkundenregesten von 54 geistlichen Institutionen[19].

Bereits in seinem ersten Arbeitsjahr 1913 begann Hoogeweg mit der Erstellung von Regesten zum Domstift Kammin. Der junge Archivar Heinrich Otto Meisner wiederum befasste sich mit der Bearbeitung der im Staatsarchiv Stettin befindlichen Unterlagen des Marienstifts Stettin[20]. Seine Aufgaben bei der Erstellung der Regesten hat kurzzeitig Paul Oberländer übernommen[21].

Anschließend (1914) begann Hermann Hoogeweg die Arbeit an den Regesten zu Urkunden des Bestandes Bistum Kammin. Er setzte diese Tätigkeit bis Ende 1916 fort und kehrte 1921–1922 noch einmal zu diesem Bestand zurück und nahm relevante Ergänzungen daran vor. Das Repertorium für Originalurkunden und Abschriften des Bistums Kammin bestand insgesamt aus 1568 Regesten in zwei Bänden sowie aus je einem Personen- und Ortsregister. Die Indizes enthielten 20.000. Einträge auf 362 Seiten[22].

---

17 H. Hoogeweg, *Die Stifter und Klöster…*, Bd. 1, S. V.

18 AP Szczecin, Sig. 65/78/0/1.1/38, S. 136 (Bericht für 1917).

19 Adolf Diestelkamp (Direktor des Stettiner Archiv im 1938) hat festgehalten, dass Hermann Hoogeweg seinerzeit 37 Repertorienbände zu allen geistlichen Urkundenbeständen angefertigt hat. A. Diestelkamp, *Das Staatsarchiv Stettin seit dem Weltkrieg*, „Monatsblätter der Gesellschaft für pommersche Gesichte und Alterthumskunde" Jg. 52 (1938), Nr 4, S. 79; M. Szukała, op.cit., s. 38; Heute sind im Stettiner Archiv noch 34 Regestenbücher erhalten: AP Szczecin, Sig. 65/78/0/3.1/1121–1153, 65/78/0/3.47/1544.

20 Ibidem, Sig. 65/78/0/1.1/38, S. 76 (Bericht für 1914).

21 Dr. Paul Oberländer wurde im Staatsarchiv Stettin am 1. April 1914 beschäftigt und nach 4 Monaten, im August wurde zur Armee eingezogen. Er starb im 1915 im Westfront.

22 Bis die Arbeit von H. Hoogeweg der Bestand Bistum Kammin erreicht nur 848 Urkunden. Ibidem, S. 76 (Bericht für 1914), 116 (Bericht für 1916).

Im Jahre 1915 beschäftigte sich Hoogeweg neben den bischöflichen Urkunden auch mit den Regesten zu Urkunden der Klöster des Regierungsbezirks Köslin. Im Jahr 1916 schloss er seine Arbeiten am Zisterzienser Nonnenkloster in Marienfließ ab. Er beendete die Bestandsaufnahme am 27. Februar – diese umfasste 27 Abschriften von Dokumenten, die sich in der Klostermatrikel in der Sammlung von Samuel Gottlieb Loeper befanden. Bis zum 1. August 1917 erstellte er die Register zu diesen Regesten. Zudem erstellte er in den Jahren 1915–1916 die Regesten für folgende Bestandsbildner: Zisterzienser Kloster Buckow (110 Urkunden), Benediktiner Nonnenkloster Altstadt Kolberg (49 Urkunden), Zisterzienser Nonnen-Kloster in Köslin (294 Urkunden), Kartause Marienkrone bei Rügenwalde (62 Urkunden)[23].

Ein Teil dieser Arbeiten war die Erweiterung der 1896 von Dr. Max Bär erstellten Urkundeninventare. Dieser hatte bereits Kurzregesten und Personenregister zu Klosterurkunden aus Buckow und Kammin erstellt[24].

Im Jahr 1917 beschäftigte sich Hermann Hoogeweg mit der weiteren Erschließung der Urkunden der Bestände Marienstift und Ottostift, die durch Dr. Paul Oberländer unvollendet abgebrochen wurde, da dieser 1914 zur Armee einberufen worden war. Hoogeweg fertigte 73 Regesten an. Die beiden Bestände umfassten nun 102 und 35 Dokumente[25]. Anschließend befasste er sich mit der Regestierung folgender Bestände: Zisterzienser Nonnenkloster vor Stettin (49 Urkunden), Kartäuserkloster Gottesgnade vor Stettin (163 Urkunden) und Jacobipriorat in Stettin (23 Urkunden). Im Jahr 1917 erstellte Hoogeweg zudem Regesten zu weiteren Beständen: Dominikaner Stolp (10 Urkunden), Prämonstratenser Kloster Belbuck (143 Urkunden), Augustiner Eremitinnen von Pyritz (35 Urkunden), Kloster der regulierten Chorherrn in Jasenitz (58 Urkunden), Kartause Gottesfriede vor Schivelbein (39 Urkunden), Zisterzienser Nonnenkloster in Wollin (45 Urkunden), Zisterzienser-Nonnenkloster Krummin (66 Urkunden), Franziskaner Kloster in Greifenberg (9 Urkunden)[26].

An der Wende von 1917 zu 1918 arbeitete er an zwei Beständen. In dieser Zeit erstellte er 377 Regesten zu Urkunden des Prämonstratenser Klosters Pudagla, sowie zu Urkunden des Domstifts Kolberg. Letztere wurden auf losen Blättern festgehalten, die dann vom Amtsgehilfen Wolter in das Regestenbuch

23 Ibidem, S. 116 (Bericht für 1916).
24 Ibidem.
25 Ibidem, S. 136 (Bericht für 1917).
26 Ibidem.

eingetragen worden sind. In dieses Buch wurden ebenfalls die Regesten für das Benediktiner Nonnenkloster Altstadt Kolberg einbezogen. Das Repertorium umfasste damit insgesamt 302 Seiten mit Registern und wurde im Juni 1918 vollendet[27]. In diesem Jahr erstellte Direktor Hoogeweg auch Regesten und Indizes für die Bestände Benediktiner Kloster Stolpe an der Peene (169 Urkunden), Benediktiner Nonnenkloster Verchen (326 Urkunden), Kloster Sankt Annen und Brigitten Marienkrone bei Stralsund (20 Urkunden)[28]. Darüber hinaus verfasste Hoogeweg in diesem Jahr Beschreibungen zu den Urkunden aus dem Depositum des Marienstifts, das vom Direktor Carl Fredrich übergeben wurde und eine Fortsetzung der schon im 19. Jh. vorhandenen Sammlung von diplomatischen Überlieferungen des Marienkapitels darstellt[29]. Die Verzeichnung dieses letzten Bestandes des Marienstiftes (Staatlicher Teil) dauerte jedoch noch bis 1920, als die Register für die 166 Regesten entstanden[30].

Im Jahr 1918 begann Hoogeweg auch mit der Inventarisierung des Urkundenbestandes des Zisterzienser Klosters Hiddensee. Bis zum Ende des Jahres hatte er 163 Regesten gefertigt, und im folgenden Jahr – weitere 200 Beschreibungen. Am Ende umfasste dieser Bestand insgesamt 403 Dokumente. 1919 erstellte er auch Regesten zu folgenden Beständen: Zisterzienser Kloster Neuenkamp (199 Urkunden), Zisterzienser Nonnenkloster zu Bergen auf Rügen (232 Urkunden) und Zisterzienser Kloster Eldena (264 Urkunden). Er vervollständigte auch die Arbeit an den Regesten zum Depositum des Marienstifts und zu den Beständen Marienstift und Ottostift, für die er Personen- und Ortsregister anfertigte[31].

Im Jahr 1920 teilte Hermann Hoogeweg zwei Dokumentensammlungen: Allgemeine geistliche Urkunden (205 Schriftstücke) und Stadt Greifswald (252 Schriftstücke). Nach dem Provenienzprinzip wurden die daraus ausgesonderten Archivalien den ursprünglichen Beständen zugeordnet. Bei den Greifswalder Urkunden ergab die Aufteilung sieben Bestände: Dominikanerkloster Greifswald (5 AE), Franziskanerkloster Greifswald (3 AE), Stankt Georgskirche Greifswald (7 AE), Heilige Geistkirche Greifswald (14/15 AE), Jakobikirche Greifswald (32/34 AE), Marienkirche Greifswald (45 AE) und Nikolaikirche und Nikolaistift

27 Ibidem, S. 151 (Bericht für 1918).
28 Ibidem.
29 Ibidem.
30 Ibidem, S. 180 (Bericht für 1920).
31 Ibidem, S. 166 (Bericht für 1919).

Greifswald (110/121 AE)[32]. Für diese Bestände transkribierte Hoogeweg 1922/1923 die erstellten Regesten in ein Findbuch. Dieses Buch mit den Regesten zu den Urkunden der Klöster und Kirchen in Greifswald wurde am 5. Januar 1923 vollendet[33].

Zudem erstellte er im Jahr 1920 die Regesten zu 8 Urkunden des Klosters Ivenack und zu 65 Urkunden des Klosters Dargun, die aus dem Bestand Allgemeine geistliche Urkunden entnommen worden waren. Da sich diese Archivalien als Urkunden des Bistums Kammin erwiesen, wurden sie von ihm – entsprechend dem Provenienzprinzip – dessen Bestand zugeordnet. Anschließend nahm er Arbeiten am Bestand Zisterzienser Nonnenklosters vor Stettin, der sich von 49 auf 146 Urkunden vergrößerte, und am Bestand des Jacobipriorats Stettin, dessen Umfang von 23 auf 164 Urkunden anwuchs, wieder vor. Die Arbeiten an den letztgenannten Urkunden setzte er noch 1921 fort, als er zusätzlich zu den Regesten auch die entsprechenden Register erstellte. Im Jahre 1920 bearbeitete er zwei kleine Sammlungen geistlicher Urkunden der Stadt Barth und der Stadt Grimmen. Nachdem er sie abgegrenzt und ihre Provenienz festgestellt hatte, bildete er zwei Bestände: Kirche der Heiligen Maria in Barth und Kirche der Heiligen Maria in Grimmen[34].

Noch 1920 führte er Arbeiten an den 1917 begonnenen Regesten für das Kloster der regulierten Chorherren in Jasenitz fort. Er schloss diese im Jahr 1921 ab und fügte den Regesten auch Personen- und Ortsregister hinzu.

Im Jahr 1921 sammelte der Archivar die bisher sehr verstreuten Unterlagen des Dominikanerklosters in Kammin (Pommern) und bildete einen gesonderten Bestand. Er erstellte für sie auch entsprechende Regesten und Register. Im gleichen Jahr schrieb er für den Bestand Bistum Kammin einen weiteren Band mit Ergänzungen, d.h. Regesten von bis dato verstreuten Urschriften zusammen mit Auszügen aus alten Repertorien. Hoogeweg beendete diese Arbeiten 1922[35].

Noch in demselben Jahr befasste er sich mit den verstreuten Archivalien des Templerordens, Johanniterordens und Deutschen Ordens, die er entsprechend ihren Provenienzen in Bestände zusammenfasste und für die er anschließend

---

32 Ibidem, S. 180 (Bericht für 1920).

33 Ibidem, Sig. 65/78/0/3.1/1136, S. 151.

34 Ibidem, Sig. 65/78/0/1.1/38, S. 180 (Bericht für 1920).

35 Ibidem, S. 190 (Bericht für 1921), 204 (Bericht für 1922).

Regesten samt Register erstellte[36]. Die letzten von Hermann Hoogeweg bearbeiteten geistlichen Urkunden betrafen das Annen- und Brigittenkloster Marienkrone bei Stralsund, sowie das Augustinerinnenkloster Armen Stralsund und das Dominikanerkloster Stralsund und das Franziskanerkloster Stralsund. Er ergänzte die bearbeiteten Regesten mit den vom Stadtarchiv Stralsund erfassten Materialien des Brigittenklosters. Endergebnis dieser Arbeiten waren folgende Bestände: das bereits erwähnte Sankt Annen und Brigittenkloster Marienkrone bei Stralsund, Oberster Kirchherr in Stralsund, Dominikaner Kloster in Stralsund, Franziskaner Kloster in Stralsund, Sankt Georg vor der Stadt, Heiliger Geist in Stralsund, Jacobikirche in Stralsund, der Kaland, Marienkirche in Stralsund sowie Nikolaikirche in Stralsund. Für diese Bestände erstellte er auch Register[37]. Die Arbeit an Urkunden, Registern und Verzeichnissen für die nachfolgenden Inventarbände setzte er bis 1923 fort. Aus diesem Grund verlängerte der Generaldirektor des Preußischen Staatsarchivs Paul Kehr – obwohl Direktor Hermann Hoogeweg im Mai 1922 das Rentenalter erreichte – sein Arbeitsverhältnis um 16 Monate, d.h. bis zum 1. Oktober 1923[38].

Die bearbeiteten Regestenbücher mit Registern betreffen, wie oben erwähnt, 55 kirchliche Anstalten, hauptsächlich Klöster in Pommern. In diesen 34 Büchern hat Herrmann Hoogeweg die Regesten für 7346 Dokumente einschließlich 3813 Originalurkunden zusammengestellt[39].

Bei der Erarbeitung dieser Findhilfsmittel hat Hoogeweg versucht, ideale Repertorien zu erstellen. Sie umfassten nicht nur Beschreibungen von Dokumenten, die im Original im Stettiner Archiv erhalten waren und die betroffenen Bestände bildeten, sondern auch Beschreibungen von Abschriften. Letztere stammen aus zwei Quellen. Das konnten zum einen Transsumpte sein, die in anderen Urkunden enthalten waren. Zum anderen handelte es sich um Abschriften von Dokumenten, die in Kopialbüchern bzw. klösterlichen, kirchlichen oder anderen (z.B. städtischen) Matrikeln gesammelt wurden. Das von Hoogeweg zusammengestellte Repertorium der Urkunden des Franziskaner

---

36 Die Urkunde der Ritterorden hat Hoogeweg auf zwei Bestände geteilt: Templer und Johanniterorden, und Deutscher Orden. AP Szczecin, Sig. 65/78/0/3.1/1141 (Regesten zu den Urkunden betreffend die Ritterorden in Pommern).

37 Ibidem, Sig. 65/78/0/1.1/38, S. 204 (Bericht für 1922).

38 Ibidem, S. 220 (Bericht für 1923); M. Szukała, op.cit., S. 60.

39 AP Szczecin, Sig. 65/78/0/3.1/1120.

Klosters in Greifenberg besteht aus nur 5 Regesten, die aus Abschriften dieser im Stadtbuch Greifenberg gesammelten Schriftstücke erstellt wurden[40]. Eine weitere Quelle für Hoogeweg waren gedruckte Urkundeneditionen und andere Publikationen. Die Regesten zum Augustiner Eremiten Kloster Marientron bei Neustettin enthalten 3 archivalische Beschreibungen: die erste ist das Regest einer in der Stettiner Monographie von Karl Tümpel veröffentlichten Urkunde (S. 42n), die zweite betrifft ein Schriftstück aus Riedls Codex diplomaticus Brandenburgensi (I, 18, S. 426, Nr. 80), und erst die dritte stellt die Beschreibung einer tatsächlich im Original erhaltenen Urkunde in dem besprochenen Bestand dar (Orig. Nr. 1)[41]. In den Jahren 1920–1921 ergänzte man die Repertorien der Bestände der Klöster in Bergen, Eldena, Neuenkamp sowie Jasenitz und Jakobipriorat Stettin durch Regesten der im Pommerschen Urkundenbuch (Bd. 1–6) bereits edierten Urkunden, die im Stettiner Archiv weder im Original noch in Abschrift erhalten geblieben sind[42]. Die Angaben zu den Quellen der Regesten (Originale, Kopialbuch, Druck) wurden immer auf der Titelseite jedes Regestenbandes angebracht.

Hermann Hoogeweg erstellte die Regesten zunächst auf Karten; teilweise verwendete er dabei Inventarkarten mit Regesten, die bereits von früheren Stettiner Archivaren, wie z.B. Max Bär 1898, zusammengestellt worden waren. Anschließend transkribierte er sie nach chronologischer Systematisierung der Dokumentenbeschreibungen in die Regestenbücher. Nachdem die Urkundenregesten in das Buch eingetragen wurden, ergänzte er es mit Personen- und Ortsregistern. Der Prozess der Endbearbeitung der Regesten und der Ergänzung der Sammlungen von Urschriften einzelner Bestände mit erhaltenen Abschriften verlorener Urkunden bewirkte, dass die Reinschrift des Findbuches mit einer Verzögerung von mehr als einem Jahr entstand. Die 1916 gefertigten Beschreibungen der Unterlagen der Kösliner Zisterzienserinnen wurden erst am 18. November 1918 von Direktor Hoogeweg in das Regestenbuch dieses Bestandes übertragen[43]. Die Reinschrift des Regestenbuches für die Urkunden

40 Ibidem, Sig. 65/78/0/3.1/1135, S. 1–3. Stadtbuch von Greifenberg ist im Bestand Akta miasta Gryfice (Magistrat Greifenberg) aufbewahrt. AP Szczecin, 65/198/0 Akta miasta Gryfice (Magistrat Greifenberg) 1501–1944, Sig. 65/198/0/1/1.

41 Ibidem, Sig. 65/78/0/3.1/1135, S. 6–9.

42 Ibidem, Sig. 65/78/0/1.1/38, S. 180 (Bericht für 1920), 190 (Bericht für 1921).

43 Ibidem, Sig. 65/78/0/3.1/1132, S. 209 (Hg. 18/11 1918, Am Tage des Einzuges der Franzosen in Strassburg).

des Klosters Kartause Marienkrone bei Rügenwalde, die im Jahr 1916 vorläufig bearbeitet wurden, vollendete er dagegen erst am 23. Oktober 1923[44].

Die von Hermann Hoogeweg erstellten Repertorien der urkundlichen Überlieferungen pommerscher Kircheninstitutionen aus dem Mittelalter sind eine unschätzbare archivarische und wissenschaftliche Hilfe für Forscher. Seit 100 Jahren werden die Regestenbücher sowohl von professionellen Historikern als auch Heimatforschern bei ihren Arbeiten genutzt. Aufgrund der Ausführlichkeit der in den Regesten enthaltenen Informationen konnten viele Forscher das Werk von Hermann Hoogeweg als Quellenmaterial für ihre Forschungen in Anspruch nehmen.

Direktor Hoogeweg selbst nutzte die Ergebnisse seiner Archivarbeit, um eine monumentale Sammelmonographie über die pommerschen Klöster sowie Dom- und Stiftskapitel zu erstellen. In dem zweibändigen Werk mit mehr als 1800 Seiten hat er die Geschichte von 53 kirchlichen Institutionen vom Gründungszeitpunkt bis zu ihrer Säkularisation während der Reformation in der ersten Hälfte des 16. Jh. zusammengestellt[45].

Die Regestenbücher sind jedoch vor allem ein archivarisches Hilfsmittel zu den Urkunden pommerscher Kirchen, die während des Zweiten Weltkrieges aus Stettin evakuiert wurden und nach 1945 in den Bestand des Landesarchivs Greifswald eingegangen sind. Im Rahmen der deutsch-polnischen Zusammenarbeit übergab das Staatsarchiv Stettin im Jahr 1969 Kopien der erhaltenen Regesten an das Landesarchiv Greifswald. Nach 50 Jahren – also im 21. Jahrhundert – stellt das Stettiner Archiv zusammen mit seinen Partnern aus Greifswald die Archivrepertorien pommerscher Kirchenurkunden in einer neuen, modernen Form bereit – elektronisch im Internet.

Der Workflow – Vom Digitalisat zum Volltext

Als technische Infrastruktur für die Volltexterkennung der handschriftlichen Urkundenregister der pommerschen Kirchen und Klöster von Hermann Hoogeweg wird die Transkribus-Plattform[46] genutzt, die im Rahmen des EU-

[44] Ibidem, Sig. 65/78/0/3.1/1142, S. 47 (Hg. 14/10 23).

[45] H. Hoogeweg, *Die Stifter und Klöster*….

[46] READ-COOP, https://readcoop.eu/ [aufgerufen am 24.6.2021]. Zur technischen Erläuterung P. Kahle, S. Colutto, G. Hackl, G. Mühlberger, *Transkribus – a Service Platform for*

Projektes READ[47] entwickelt wurde. Der Workflow[48] für die Verarbeitung der Handschriften auf der Transkribus-Plattform besteht im Wesentlichen aus 5 Schritten:

a) Digitalisate einspielen

Um eine automatisierte Verarbeitung der Dokumente durchführen zu können, müssen die Digitalisate zuerst auf die Transkribus-Plattform hochgeladen werden. Dabei wird eine eigene Sammlung („collection") angelegt. Jeder Band der Regesten wird dabei als ein eigenes Dokument deklariert. Wer auf die entsprechenden Sammlung zugreifen kann, wird von einem Administrator festgelegt, der den einzelnen Nutzern auch unterschiedlich weit gehende Rechte und Rollen zuweisen kann. Fremde Personen haben keine Zugriffe auf das Material.

Die von uns verwendeten Digitalisate, die im Staatsarchiv Stettin angefertigt wurden, weisen eine Auflösung von 400 dpi auf. Diese Auflösung ist für die folgende Verarbeitung optimal. Es können in Transkribus aber grundsätzlich auch Digitalisate mit geringeren Auflösungen erfolgreich verarbeitet werden.

b) Segmentierung

Anschließend werden die Digitalisate einer automatischen Layoutanalyse unterzogen, d.h. die Bereiche jeder Seite, auf denen sich Text befindet, werden automatisch markiert und segmentiert und eine Zeilenerkennung durchgeführt. HTR ist – anders als OCR, wo eine Erkennung auf der Ebene von Einzelbuchstaben und Wörtern erfolgt – ein zeilenbasiertes Verfahren[49]. Daher ist eine weitere Segmentierung des Materials nicht nötig.

Im Falle der Regestenbücher, die in einer tabellarischen Form aufgebaut sind, haben wir uns für eine spezielle Art der Segmentierung entschieden.

Transcription, Recognition and Retrieval of Historical Documents, https://www.researchgate.net/publication/322780398_Transkribus_-_A_Service_Platform_for_Transcription_Recognition_and_Retrieval_of_Historical_Documents [aufgerufen am 24.6.2021].

47 G. Mühlberger, L. Seaward, M. Terras et al., *Transforming scholarship in the archives through handwritten text recognition. Transkribus as a case study*, „Journal of Documentation" 75/5 (2019), S. 954–976, hier S. 957f., https://www.research.ed.ac.uk/en/publications/transforming-scholarship-in-the-archives-through-handwritten-text [aufgerufen am 24.6.2021].

48 Ibidem, S. 958–964.

49 G. Mühlberger, *Archiv 4.0 oder warum die automatisierte Texterkennung alles verändern wird*, w: *Massenakten – Massendaten. Rationalisierung und Automatisierung im Archiv. 87. Deutscher Archivtag 2017 in Wolfsburg (Tagungsdokumentationen zum Deutschen Archivtag, Bd. 22)*, Hrsg. von K. Deecke, E. Grothe, Fulda 2018, 145–156.

Wir verwenden dafür ein Tabellenlayout[50], das die einzelnen Spalten der Regesteneinträge, wie „laufende Nummer“, „Nummer des Originals“, „Datum“ und „Regest“ voneinander trennt und zusätzlich jedes einzelne Urkundenregest vom anderen abgrenzt. Innerhalb dieser Tabellenstruktur werden dann die einzelnen Zeilen automatisch erkannt.

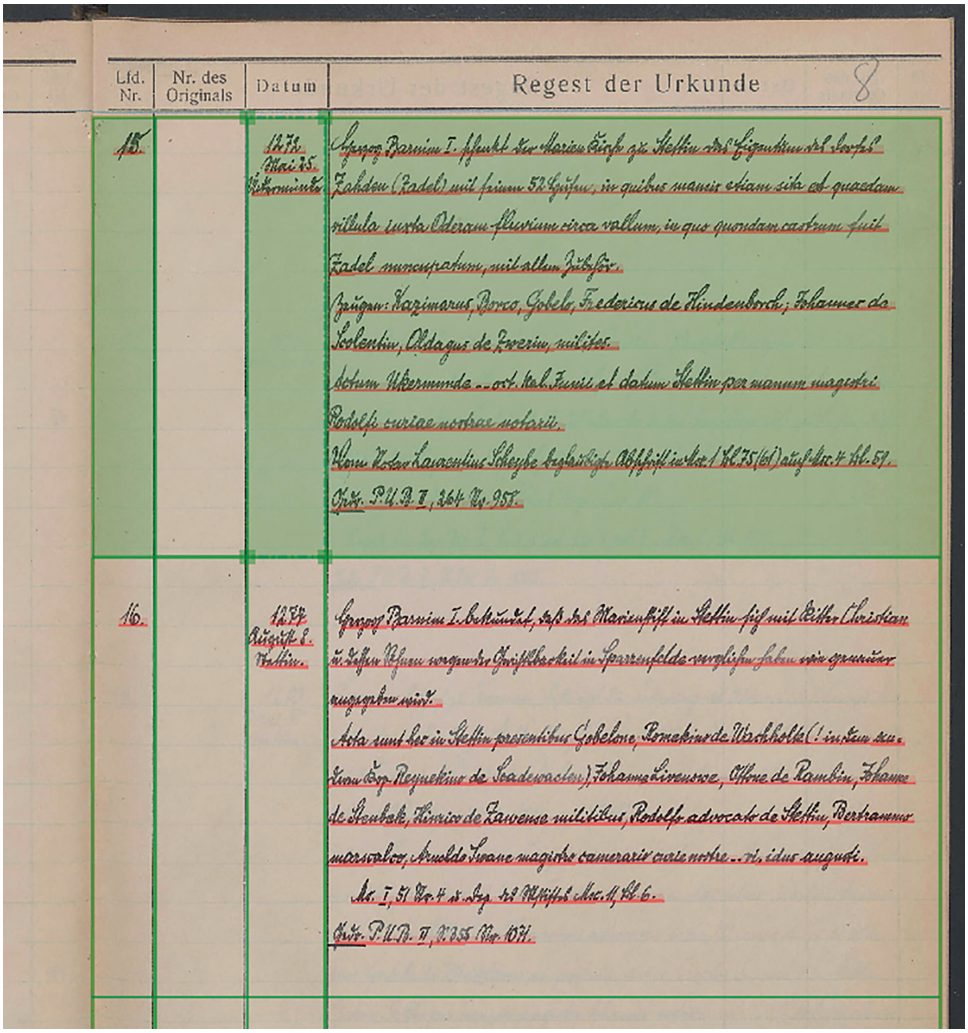


Abb: Seite aus den „Regesten zu dem vom Marienstift Stettin deponierten Urkunden“, bearbeitet von Meiser, Oberländer und Hoogeweg 1913–1917 (AP Szczecin, Sig. 65/78/0/3.47/1544) nach der Segmentierung in Transkribus.

[50] Zur Funktionsweise vgl. READ COOP, Wie man mit Tabellen in Transkribus arbeitet, https://readcoop.eu/transkribus/howto/how-to-work-with-tables-in-transkribus/ [aufgerufen 24.6.2021].

Die Vorzüge dieses „table editing" bestehen vor allem darin, dass die späteren Regestendaten übersichtlich geordnet präsentiert werden können. Darüber hinaus besteht bei dieser Art der Segmentierung die Möglichkeit, die „Leseergebnisse" nicht nur als Fließtext, sondern auch als Datentabelle zu exportieren[51]. Dadurch wiederum wird eine Nachnutzung wesentlich erleichtert.

c) Transkription

Das Kernstück jeder Handschriftentexterkennung sind die dabei verwendeten HTR-Modelle. Dabei werden Verfahren des „Deep Learning" eingesetzt, um den neuronalen Netzen das „Lesen zu lernen". Inzwischen gibt es bereits fast 100 frei und kostenlos verfügbare HTR-Modelle, die von jedermann genutzt werden können – darunter HTR-Modelle für deutsche Kurrentschrift, „dutch writing" oder auch moderne polnische Handschriften[52]. Dabei handelt es sich zumeist um sogenannte generische Modelle, die ein Vielzahl von Anwendungsfällen abdecken sollen.

Angesichtes der begrenzten Anzahl von Schreibern in unserem Projekt, haben wir uns dafür entschieden, ein eigenes Spezialmodell zu entwickeln, das diese Handschriften in der größtmöglichen Qualität automatisch entziffert. Dafür ist es nötig, dass zunächst eine gewisse Menge Seiten der drei Handschriften (von Meisner, Oberländer und Hoogeweg), in denen die Regestenbände aufgezeichnet sind fehlerfrei transkribiert werden. Diese Arbeit wurde im Universitätsarchiv Greifswald und an der Universität Kiel erledigt. Diese Seiten dienen dann als zuverlässiges Trainingsmaterial für die HTR-Modelle und werden als „Ground Truth" bezeichnet. Als Faustregel wird häufig angegeben, dass man 20.000–30.000 Wörter Ground Truth (das sind etwa 100 bis 150 Textseiten) benötigt, um ein leistungsfähiges HTR-Modell zu entwickeln[53]. Die nötigen Transkriptionen können komfortabel und manuell im Expert Client von Transkribus angefertigt werden, wobei die jeweilige Transkription der entsprechenden Zeile im Digitalisat zugeordnet wird[54]. Soweit vorhanden, lassen sich aber auch bereits existierende

---

51 Zu den Exportfunktionen von Tanskribus vgl. READ COOP, So exportieren Sie Dokumente aus Transkribus, https://readcoop.eu/transkribus/howto/how-to-export-documents-from-transkribus/ [aufgerufen 24.6.2021].

52 READ COOP, Öffentliche AI-Modelle in Transkribus, https://readcoop.eu/transkribus/public-models/ [abgerufen am 5.7.2021].

53 Vgl. G. Mühlberger, L. Seaward, M. Terras et al., op.cit., S. 959; G. Mühlberger, T. Terbul, *Handschriftenerkennung für historische Schriften. Die Transkribus Plattform*, „b.i.t. online. Bibliothek. Information. Technologie" Jg. 21 (2018) Nr. 3, S. 218–222.

54 Eine Einführung dazu findet man hier: READ COOP, Wie man Dokumente mit Transkribus transkribiert – Einführung, https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/ [aufgerufen am 24.6.2021].

Transkriptionen, z.B. aus Editionen oder Abschriften verwenden und somit „recyceln“[55].

Für das Training eines HTR-Modells, das die drei Handschriften der Stettiner Regestenbände problemlos lesen kann, haben wir zunächst mit einem Ground Truth von 30.0000 Wörtern begonnen, was bereits zu sehr guten Ergebnissen führte. Derzeit besteht unser Trainingsmaterial aus etwa 100.000 Wörtern.

d) Training des HTR-Modells

Der unbestreitbare Vorteil der Transkribus-Plattform ist, dass sie (als bisher einzige) den freien Zugang zum Training eigener und spezifischer HTR-Modelle für jedermann ermöglicht. Dadurch kann man für jeden beliebigen Anwendungsfall ein Texterkennungsmodell entwickeln, das der jeweiligen Schrift, dem Sprachstand und den eigenen Transkriptionsrichtlinien angepasst ist[56].

Diese HTR-Modelle sind Repräsentationen des „Wissens“, das beim Trainingsprozess von der Maschine selbständig angeeignet wird. Das Wissen der neuronalen Netze bzw. der Modelle beruht jedoch immer auf reiner Statistik. Je regelmäßiger eine Schrift und je höher der Informationsgehalt einer Schrift, desto erfolgreicher wird der Lernprozess sein. Auf Grundlage der Trainingsdaten treffen dann die Modelle bei der eigentlichen Erkennung Entscheidungen, die auf Wahrscheinlichkeiten beruhen. Das geht soweit, dass für jeden einzelnen Buchstaben in einer Zeile eine Wahrscheinlichkeitsmatrix existiert, aus der dann – basierend auf weiteren Informationen wie etwa einem Sprachenmodell – der wahrscheinlichste Buchstabe ausgewählt wird.

Für die Durchführung der HTR stehen in Transkribus zwei HTR-Programme („engines“) zur Verfügung – „HTR+“ und „PyLaia“. Im Regestenprojekt haben wir zunächst mit dem Training von HTR+-Modellen begonnen, weil dieses Programm im Anfangsstadium des Projektes (wenn noch wenig Ground Truth für das Training zur Verfügung steht) schneller „lernt“ und zuverlässiger „liest“.

Für unser Projekt waren beim Training eines HTR-Modells verschiedene Aspekte zu berücksichtigen. Die Regestenbände sind in deutscher Kurrentschrift

---

55 READ COOP, Verwendung vorhandener Transkriptionen zum Trainieren eines HTR-Modells mit dem TextToImage-Tool, https://readcoop.eu/transkribus/howto/how-to-use-existing-transcriptions-to-train-a-handwritten-text-recognition-model/ [aufgerufen am 24.6.2021].

56 Vgl. mit vielen Beschreibungen und Anregungen zum HTR-Training: A. Rabus, *Trainig generic models for Handwritten Text Recognition using Transkribus: Oppotunities and pitfalls*, https://www.academia.edu/49356690/Training_generic_models_for_Handwritten_Text_Recognition_using_Transkribus_Opportunities_and_pitfalls [abgerufen am 25.6.2021].

von drei verschiedenen Schreibern verfasst worden (Hoogeweg, Oberländer, Meisner). Das HTR-Modell sollte also in der Lage sein, alle drei Handschriften möglichst gleich gut zu lesen. Hinzu kam, dass wir es mit drei verschiedenen Sprachen in den Regesten zu tun haben: Deutsch, Mittelniederdeutsch und Latein.

Um ein gutes Ergebnis zu erzielen musste aus den Regestenbänden also ein repräsentativer Querschnitt von Seiten ausgewählt werden, die sowohl die verschiedenen Schreiberhände als auch die unterschiedlichen Sprachen exemplarisch abbildete, damit das HTR-Modell in der Lage war, alle Besonderheiten zu erlernen.

Der Erfolg eines HTR-Modells wird in der sogenannten Character Error Rate (CER) und der Word Error Rate (WER) bemessen. Sie geben an, wieviele Textoperationen durchgeführt werden müssen um aus einem fehlerhaften Text den korrekten Text erzeugen zu können. Es werden dabei drei Operationen unterschieden: Ersetzungen, Hinzufügungen und Löschungen. Nach mehreren Trainingsdurchgängen konnten wir beim Einsatz von ca. 100.000 Wörtern Ground Truth im Trainingsset eine durchschnittliche CER von 1,75 % am Validierungsset erzielen. Das heißt, dass die automatische Texterkennung etwa 98% der Buchstaben des gesamten Textes fehlerfrei gelesen hat.

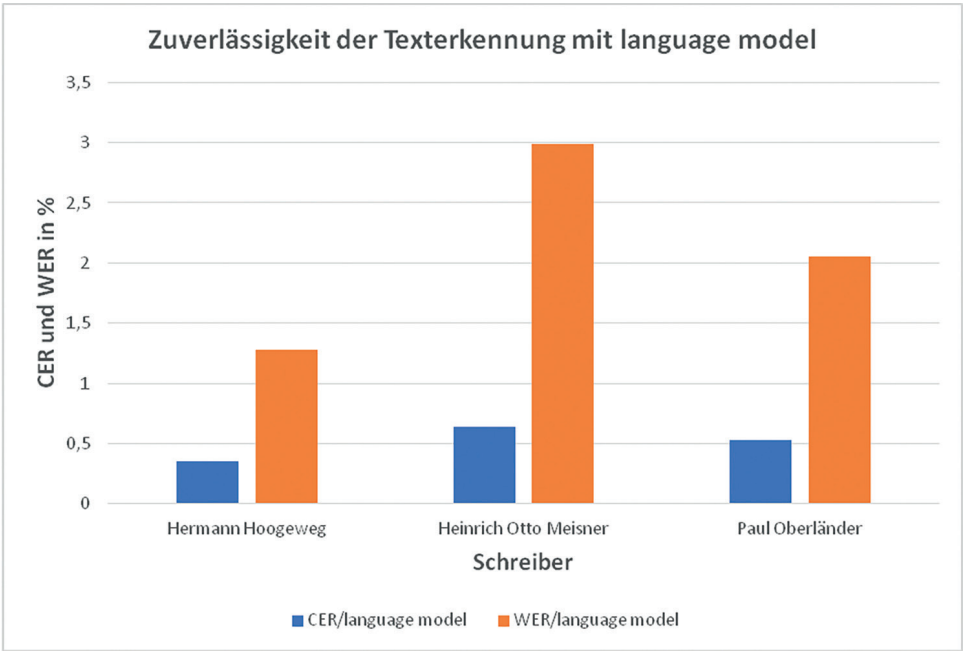
Das HTR-Modell „liest" einen Text allein aufgrund seiner grafischen Merkmale, also der Form der Zeichen und Linien. Um das Ergebnis eines solchen HTR-Modells in Transkribus noch zu verbessern, werden parallel zum „Optischen Modell" auch sogenannte „Sprachmodelle"[57] trainiert.

Diese Sprachmodelle ergänzen die „paläografische Sicht" des HTR-Modells wiederum durch eine einfache Statistik der Anordnung von Buchstaben im Trainingsset. Das führt besonders bei größeren Modellen, wie dem unseren, zu einer merklichen Verbesserung der Ergebnisse besonders was die Word Error Rate betrifft.

Betrachtet man das Leseergebnis unseres HTR-Modells mit dem zusätzlichen Einsatz des Sprachenmodells auf die individuellen Handschriften des Regestenmaterials bezogen, dann ergibt sich folgendes Bild:

---

[57] T. Strauß, M. Weidemann, R. Labahn, *Recognition and Enrichment of Archival Documents. D7.11. Language Models. Improving transcriptions by external language resource,* 2017, https://readcoop.eu/wp-content/uploads/2017/12/D7.11_final.pdf [abgerufen am 25.6.2021].



Die Grafik zeigt, dass die Handschriften der Regestenschreiber Hoogeweg, Meisner und Oberländer mit einer durchschnittlichen Genauigkeit von über 99% erkannt werden. Betrachtet man zusätzlich die Word Error Rate, dann bedeutet das, dass bei einer Volltextsuche im gesamten Regestenmaterial sicher 98% der Wörter richtig erkannt und gefunden werden können.

e) Präsentation der Ergebnisse

Um die Ergebnisse unserer Arbeit der Forschung und der breiten Öffentlichkeit zugänglich zu machen präsentieren wir die Digitalisate und die automatisch generierten Volltexte in der Digitalen Bibliothek Mecklenburg-Vorpommerns[58]. Dort stehen Suchwerkzeuge zur Verfügung, die es gestatten sich die Suchtreffer durch Highlighting auch im Digitalisat anzeigen zu lassen.

[58] Die Digitale Bibliothek Mecklenburg-Vorpommern, Regesten zu den Urkunden der pommerschen Kirchen und Klöster - Regesty dokumentów kościołów i klasztorów pomorskich, op. cit.

Außerdem kann man in einem speziellen Viewer Digitalisat und Volltext parallel betrachten:

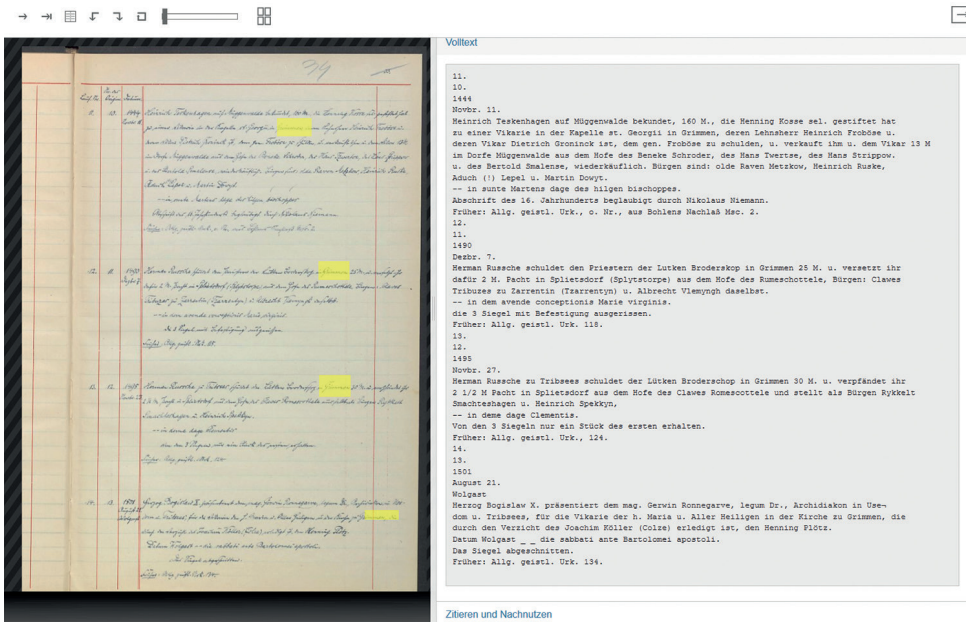


Abb: Ansicht des Viewers der Digitalen Bibliothek Mecklenburg-Vorpommern mit der Suche nach dem Begriff „Grimmen“ im Regestenband zu den Urkunden der Kirchen in Barth und Grimmen und Trefferanzeige mit Highlighting im Digitalisat (AP Szczecin, Sig. 65/78/0/3.1/1122).

Um eventuelle Verluste in der Texterkennung auszugleichen, wird in wenigen Wochen noch eine zusätzliche, spezielle Suchtechnik ermöglicht – das Keyword Spotting (KWS). Dazu wird eine eigene Website mit einem besonderen Viewer aufgesetzt, auf der diese Suchen durchgeführt werden können[59]. Anders als bei der herkömmlichen Volltextsuche, wird dabei nicht der automatisch erkannte Volltext durchsucht. Stattdessen durchsucht das Programm einen Index, der sich die oben ausgeführte Eigenschaft des Deep Learning zunutze macht, weshalb auch Wörter gefunden werden können, die bei der wahrscheinlichsten Lesung unter Umständen verworfen worden sind. Diese „verworfenen Lesungen“ werden nun beim Keyword Spotting ebenfalls durchsucht. Dadurch werden eben auch Wörter gefunden, die obwohl sie vielleicht richtig waren, vom HTR-Modell bei

[59] https://transkribus.eu/r/regestapomeraniae/#/ [ab 1.12.2021 freigeschaltet].

der Ausgabe des Volltextes nicht berücksichtigt wurden. Der besondere Vorteil des KWS liegt dabei darin, dass verschiedenen Schreibweisen von Orts- und Personennamen gefunden werden, auch wenn nicht explizit danach gesucht wird. Gerade in den Regesten sind regelmäßig die Personen, Orts- und Flurnamen aus den Urkunden in wechselnden Schreibweisen wiedergegeben worden. Und hier hilft das KWS ganz erheblich zuverlässige Suchtreffer zu erzielen. Im Regestenbuch für das Kloster Hiddensee wird der Ort „Ueckermünde" (heutige Schreibweise) bspw. als „Uckermünde" oder „Ückermünde" angegeben. Bei einer reinen Volltextsuche würde immer nur der „Ausdruck" gefunden werden, den man im Suchfeld eingibt. Mit Hilfe des KWS lassen sich bei der Eingabe der modernen Schreibweise „Ueckermünde" – die eigentlich im Text von Hoogeweg nicht vorkommt sämtliche Textstellen finden, die sich auf diesen Ort beziehen, egal ob sie als „Uckermunde" oder „Ückermünde" geschrieben werden.

Ausblick

Bis zum Ende des Jahres 2021 werden im Rahmen des polnisch-deutschen Kooperationsprojektes und unter Einsatz der Handschriftentexterkennung die Inhalte der Regesten von über 7000 pommerschen Urkunden für die wissenschaftliche Forschung in Polen und Deutschland und darüber hinaus weltweit im Internet verfügbar gemacht werden. Wir hoffen, dass dieser erfolgreichen Kooperation weitere folgen, in denen wir unser gemeinsames historische Erbe für die internationale Forschung erschließen.

Quelle

Archiwum Państwowe w Szczecinie

65/78/0 Archiwum Państwowe w Szczecinie (Staatsarchiv Stettin) [1500] 1831–1945 [1971], Sig. 65/78/0/1.1/38, 65/78/0/2/1116, 65/78/0/3.1/1120, 65/78/0/3.1/1122, 65/78/0/3.1/1132, 65/78/0/3.1/1135, 65/78/0/3.1/1136, 65/78/0/3.1/1141, 65/78/0/3.1/1142, 65/78/0/3.47/1544.

65/198/0 Akta miasta Gryfice (Magistrat Greifenberg) 1501–1944, Sig. 65/198/0/1/1.

Web-Seite

Christian-Albrechts-Universität zu Kiel, Das Klosterregister und Klosterbuch für Pommern, https://www.histsem.uni-kiel.de/de/das-institut-1/abteilungen/regionalgeschichte-mit-schwerpunkt-schleswig-holstein/projekte/pommersches-klosterbuch [abgerufen am 28.6.2021].

De Gruyter, Digitalisierungsprojekte des Staatsarchivs Zürich mit Einsatz von Machine-Learning-Verfahren, https://www.degruyter.com/document/doi/10.1515/abitech-2020-2018/html [abgerufen am 5.7.2021].

Die Digitale Bibliothek Mecklenburg-Vorpommern, Regesten zu den Urkunden der pommerschen Kirchen und Klöster - Regesty dokumentów kościołów i klasztorów pomorskich, https://www.digitale-bibliothek-mv.de/viewer/toc/PPNAPSzczecinie_65_78_0_3_1/ [abgerufen am 5.7.2021].

Gemeente Amsterdam, Vreemdelingenkaarten, https://www.amsterdam.nl/stadsarchief/nieuws/vreemdelingenkaarten/ [abgerufen am 28.06.2021].

Österreichische Nationalbibliothek, Austrian Books Online, https://www.onb.ac.at/digitaler-lesesaal/austrian-books-online-abo [abgerufen am 24.6.2021].

READ COOP, Öffentliche AI-Modelle in Transkribus, https://readcoop.eu/transkribus/public-models/ [(abgerufen am 5.7.2021]) [aufgerufen am 24.6.2021].

READ COOP, So exportieren Sie Dokumente aus Transkribus, https://readcoop.eu/transkribus/howto/how-to-export-documents-from-transkribus/ [aufgerufen 24.6.2021].

READ COOP, Verwendung vorhandener Transkriptionen zum Trainieren eines HTR-Modells mit dem TextToImage-Tool, https://readcoop.eu/transkribus/howto/how-to-use-existing-transcriptions-to-train-a-handwritten-text-recognition-model/ [aufgerufen am 24.6.2021].

READ COOP, Wie man Dokumente mit Transkribus transkribiert – Einführung, https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/ [(zuletzt aufgerufen am 24.6.2021]).

READ COOP, Wie man mit Tabellen in Transkribus arbeitet, https://readcoop.eu/transkribus/howto/how-to-work-with-tables-in-transkribus/ [aufgerufen 24.6.2021].

READ-COOP, https://readcoop.eu/ [aufgerufen am 24.6.2021].

Rechtsprechung im Ostseeraum, https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/ [abgerufen am 5.7.2021].

Search Finnish Court Records. Search and browse district court records from 1810 to 1870, https://tuomiokirjat.narc.fi/en [abgerufen am 28.6.2021].

Semantic Computing Research Group (SeCo), Handwritten Text Recognition in the Archives, https://seco.cs.aalto.fi/events/2018/2018-10-23-heldig-summit/presentations/06-kallio.pdf [abgerufen am 28.6.2021].

Transkribus, Amsterdam notarial deeds, https://transkribus.eu/r/notarial/ [(abgerufen am 28.6.2021]).

Zoeken in transcripties, www.zoekintranscripties.nl und den Vortrag von L. Keyser auf YouTube, 02 Transkribus in practise – Transkribus User Conference 02/2020, https://www.youtube.com/watch?v=xQPcJHGn8cM&t=866s [abgerufen am 28.6.2021].

Literatur

Diestelkamp A., *Das Staatsarchiv Stettin seit dem Weltkrieg*, „Monatsblätter der Gesellschaft für pommersche Gesichte und Alterthumskunde" Jg. 52 (1938), Nr 4, S. 71–82.

Hoogeweg H., *Die Stifter und Klöster der Provinz Pommern*, Bd. 1–2, Stettin 1924–1925.

Hoogeweg H., *Verzeichnis der Stifter und Klöster Niedersachsens vor der Reformation, umfassend die Provinz Hannover, die Herzogtümer Braunschweig und Oldenburg, die Fürstentümer Lippe-Detmold und Schaumburg-Lippe, die Freien Städte Bremen und Hamburg und Hessisch-Schaumburg*, Hannover 1908.

Kahle P., Colutto S., Hackl G., Mühlberger G., *Transkribus – a Service Platform for Transcription, Recognition and Retrieval of Historical Documents,* https://www.researchgate.net/publication/322780398_Transkribus_-_A_Service_Platform_for_Transcription_Recognition_and_Retrieval_of_Historical_Documents [aufgerufen am 24.6.2021].

Mühlberger G., *Archiv 4.0 oder warum die automatisierte Texterkennung alles verändern wird*, w: *Massenakten – Massendaten. Rationalisierung und Automatisierung im Archiv. 87. Deutscher Archivtag 2017 in Wolfsburg (Tagungsdokumentationen zum Deutschen Archivtag, Bd. 22)* Hrsg. Von K. Deecke, E. Grothe, Fulda 2018, 145–156.

Mühlberger G., Seaward L., Terras M. et al., *Transforming scholarship in the archives through handwritten text recognition. Transkribus as a case study*, „Journal of Documentation" 75/5 (2019), S. 954–976, hier S. 957f.

Mühlberger G., Terbul T., *Handschriftenerkennung für historische Schriften. Die Transkribus Plattform*, „b.i.t. online. Bibliothek. Information. Technologie" Jg. 21 (2018) Nr. 3, S. 218–222.

Pommersches Urkundenbuch, Bd. 1 Abt. 1, 786–1253. Bearbeitet und herausgegeben von R. Klempin, Stettin 1868.

Rabus A., *Trainig generic models for Handwritten Text Recognition using Transkribus: Oppotunities and pitfalls,* https://www.academia.edu/49356690/Training_generic_models_for_Handwritten_Text_Recognition_using_Transkribus_Opportunities_and_pitfalls [zuletzt abgerufen am 25.6.2021].

Strauß T., Weidemann M., Labahn R., *Recognition and Enrichment of Archival Documents. D7.11. Language Models. Improving transcriptions by external language resource*, 2017, https://readcoop.eu/wp-content/uploads/2017/12/D7.11_final.pdf [(zuletzt abgerufen am 25.6.2021]).

Szukała M., *Archiwum Państwowe w Szczecinie w latach 1914–1945. Ludzie i działalność*, Szczecin 2019.