Informationswissenschaft: Theorie, Methode und Praxis
ISSN: 2297-9069
gaby.knoch-mund@hist.unibe.ch
Universität Bern
Suiza

Ansätze zur computergestützten Erschliessung von gleichförmigen Massenakten

Zimmermann, Lynn
Ansätze zur computergestützten Erschliessung von gleichförmigen Massenakten
Informationswissenschaft: Theorie, Methode und Praxis, vol. 7, núm. 1, 2022
Universität Bern, Suiza
Disponible en: https://www.redalyc.org/articulo.oa?id=664372181019
DOI: https://doi.org/10.18755/iw.2022.20

Ansätze zur computergestützten Erschliessung von gleichförmigen Massenakten

Lynn Zimmermann

Informationswissenschaft: Theorie, Methode und Praxis, vol. 7, núm. 1, 2022

Universität Bern, Suiza

DOI: https://doi.org/10.18755/iw.2022.20

Redalyc: https://www.redalyc.org/articulo.oa?id=664372181019

Die vorliegende Untersuchung zeigt auf, ob und wie mithilfe von computergestützten Verfahren die Erschliessungsarbeit in Archiven erleichtert, beschleunigt und verbessert werden kann. Der Fokus liegt dabei auf gleichförmigen, analogen Massenakten des 19. und 20. Jahrhunderts, die für viele Archive aufgrund ihrer grossen Menge ein Ressourcenproblem darstellen: sowohl der physische Raum, den sie einnehmen, wie auch die Zeit, die für die archivkonforme Erschliessung benötigt wird, sind knapp bemessen. Auf der einen Seite befinden sich Staatsarchive, die darum bemüht sind, diese Massenakten zugänglich zu machen und oft Kompromisse eingehen müssen, um die ausufernden Bestände der Verwaltungsakten sach- aber auch termingerecht zu erschliessen. Auf der anderen Seite stehen Google-erprobte Benutzer und Benutzerinnen, die erwarten, dass die Suchmaske des Archivkatalogs bereits mit einem Stichwort und wenigen Mausklicks eine zufriedenstellende Trefferliste ausgibt. Ein grosses Wissen über die Verwaltungsgeschichte kann beim externen Klientel nicht vorausgesetzt werden, weshalb sowohl die Suchmaschine als auch der Archivar mithilfe einer aussagekräftigen Titelvergabe diese Wissenslücke überbrücken müssen. Wird kein Suchtreffer erzielt, dann wird der Bestand oder das Dossier nicht benutzt und ist - zumindest aus Benutzersicht - gar nicht vorhanden.

Die folgenden Seiten beschreiben, inwieweit die Erschliessungsarbeit durch Scanner, Computerprogramme und weitere technische Hilfsmittel unterstützt werden kann, sodass gegebenenfalls gar keine Titel mehr händisch vergeben werden müssen. Anhand einer Kartei über die Grossratsmitglieder des Kantons Thurgau zwischen 1803 bis 1930 wird nachvollzogen, inwieweit eine computergestützte Erschliessung dieser Karteikarten durchführbar ist. Nachdem der Bestand gescannt und ein Modell für die automatische Transkription dem Texterkennungsprogramm für Handschriften «Transkribus» antrainiert worden ist, sollen Textregionen definiert werden, die für den Dossiertitel relevant sind. Diese Textregionen werden exportiert, mithilfe eines Scripts gefiltert, in einem Texteditor nachbearbeitet und zuletzt ins Archivinformationssystem importiert. Dabei werden diese Prozesse fortlaufend überprüft und protokolliert, um die automatisierte Titelvergabe mit der

herkömmlichen Erschliessung von Hand zu vergleichen.

1. Technische und rechtliche Voraussetzungen

1.1 Die Archives nationales de France erschliessen mehr als 1800 Notariatsregister

Auf der Exkursion des Studiengangs in die *Archives nationales de France* in Paris präsentierte eine Forschergruppe ihre Arbeit im *Département du minutier central des notaires de Paris*. Sehr grosse Teile der Notariatsakten von Paris wurden als Digitalisate der Verwaltung und der Öffentlichkeit zur Verfügung gestellt [1] . Weil die Navigation in den Scans als auch den analogen Notariatsakten aufgrund des stark strukturierten Inhalts mit vielen Querverweisen zeitaufwändig ist, sollten die Digitalisate transkribiert werden. In einer Transkription könnte man mit der Volltextsuche schnell auf den gewünschten Treffer stossen und müsste weniger Vorwissen zu den verschiedenen Verwaltungsakten mitbringen, um sich in den komplexen Notariatsakten zurechtzufinden. Weil die herkömmliche Transkription von Hand wegen des sehr grossen Bestands nicht innert einer nützlichen Frist ausführbar wäre, wurde die Handschriftenerkennungssoftware Transkribus [2] beigezogen. Diese Software bietet den Vorteil, dass man nur einen Bruchteil des Korpus transkribieren muss und dabei gleichzeitig ein Modell trainiert. Dieses kann dann auf die nicht transkribierten Teile angewendet werden und liefert sehr exakte Ergebnisse [3] .

Die Software Transkribus, die innerhalb der von der EU geförderten Forschungsprojekte tranScriptorium und READ (Recognition and Enrichment of Archival Documents) entwickelt worden ist, hat das Ziel, den Zugang zu den Quellen für die Forschung zu erleichtern. Texte müssen nicht mehr vollständig durchgelesen oder transkribiert werden, um mit ihnen arbeiten zu können.

Transkribus kann auch in den Strukturen des Textes suchen. Überschriften oder Marginalien werden nach dem Trainieren eines weiteren Modells erkannt, wodurch die Navigation im Text erleichtert wird. Andererseits ist es auch möglich, mithilfe der Stichwortsuche aus einem unbekannten Manuskript gewisse besonders relevante Textteile zu extrahieren und losgelöst vom umgebenden Inhalt zu untersuchen.

Um den Vorgang der Transkription und die Aufbereitung der Notariatsakten des Pariser Beispiels nachzuvollziehen, musste im Staatsarchiv ein geeigneter Bestand gefunden werden. Davor galt es aber, die Rechtslage zu klären. Besonders personensensible Daten dürfen innerhalb der Schutzfrist nicht in unbefugte Hände geraten und müssen mit der nötigen Umsicht behandelt werden. Diese Personendaten

dürfen nur verwendet werden, wenn diejenige Person oder Stelle, die mit den Daten arbeiten will, den Datenschutz ständig gewähren kann. [4] Um diese Vorgaben einzuhalten, wäre ein grenzüberschreitendes Übereinkommen mit der Universität Innsbruck, die die Server von

Transkribus hostet, nötig gewesen. Da Transkribus auf Crowdsourcing setzt, um einen möglichst grossen Quellen- und Transkriptionskorpus anzulegen und damit präziser zu werden, werden solche Übereinkommen grundsätzlich nicht gemacht.

Zudem spielen Schutzfristen für die frei zugänglichen Modelle in Transkribus, die grössere Korpora abdecken, keine Rolle. Sie basieren einerseits auf Zeitungen und anderen Druckerzeugnissen, die seit ihrer Entstehung öffentlich zugänglich sind. Andererseits basieren sie auf Handschriften, die wegen ihres hohen Alters keiner Schutzfrist mehr unterstehen.

Aufgrund der gesetzlichen Vorgaben des Datenschutzgesetzes des Kantons Thurgau und der Nutzungsbestimmungen von Transkribus musste für den Testlauf darum ein Bestand gefunden werden, der entweder keine Personendaten enthält oder dessen personenbezogene Daten keiner Schutzfrist mehr unterstehen oder nie unterstanden haben, da die Betroffenen Personen des öffentlichen Interesses waren.

1.2 Umgang mit seriellen Massenakten

Texterkennungsprogramme werden in vielen Scannern bereits ab Werk mitgeliefert. Die meisten dieser Geräte arbeiten mit OCR (Optical Character Recognition). Mit ihr können PDFs und verwandte Formate bearbeitet werden. Sie arbeitet grundlegend folgendermassen: Ein Scanner oder eine Kamera fertigen beim Einlesen des Dokuments eine Rastergrafik an. Diese Grafik besteht aus hellen und dunklen Pixeln. Ein Texterkennungsprogramm mit OCR versucht nun, die dunklen Pixel, die nahe beieinander sind, mit einem vorhandenen Modellzeichensatz abzugleichen und das Zeichen mit den grössten Gemeinsamkeiten zu finden. Der Scan wird Zeichen für Zeichen ausgelesen und jeweils einem Buchstaben, Zahlenwert oder Sonderzeichen zugeordnet. Weil die OCR allerdings nur einzelne Zeichen, nicht aber zusammenhängende Wörter, das heisst Sinneinheiten erkennen kann, kann es insbesondere bei schlechteren Scans zu fehlerhaften Resultaten kommen [5].

Die OCR basiert auf einem statistischen Zeichensatz. Der zu erkennende Text wird mithilfe eines Algorithmus Zeichen für Zeichen abgesucht und mit einem Thesaurus abgeglichen. Kommt ein Zeichen des Textes nicht im Modell der hinterlegten Zeichen vor oder weicht zu stark davon ab, wird es mithilfe eines

Annäherungsverfahrens bestimmt. Das in Form und Grösse ähnlichste Zeichen wird dann als richtig angenommen [6].

Für moderne Schriftarten ist die OCR mittlerweile so weit fortgeschritten, dass die Zielsetzung, eine genaue und hohen Ansprüchen genügende Zeichenerkennung zu entwickeln, bereits erreicht ist. Versucht man aber, diese Modelle auf ältere Druckschriften oder sogar Handschriften anzuwenden, kommt die OCR an ihre Grenzen.

Statt wie bei der OCR-Erkennung von einem Masterzeichensatz auszugehen und diesen Buchstabe für Buchstabe dem auszulesenden Text überzustülpen, basieren moderne Texterkennungsprogramme auf

Methoden, die einander ganze Zeilen gegenüberstellen. [7] Die Zeilen werden identifiziert, indem ein Algorithmus den Kontrast eines Pixels mit dem Kontrast der angrenzenden Pixel vergleicht. Ist der Kontrast in einem Pixelumfeld gleichbleibend dunkel oder schwarz, dann erkennt der Algorithmus diese Pixelabfolge als Zeile. Alle Abschnitte mit Pixeln, die hell oder weiss erscheinen, sind demzufolge Zwischenräume zwischen den Zeilen. Alle Abschnitte mit dunklen oder schwarzen Pixeln müssen im Idealfall Teil eines Zeichens sein. Dieses sogenannte «seam carving», das den bedeutungstragenden Raum in einem Scan vom nicht bedeutungstragenden Teil absondert, sagt noch nichts über den Inhalt des Textes aus. Die Menge der dunklen und schwarzen Pixel lässt noch nicht erahnen, ob es sich beim Bildpunkt um einen Teil eines Buchstabens, eines Tabellenstrichs oder um ein Staubkorn handelt. Gleichzeitig werden auch hinterlegte Layoutdaten miteinbezogen [8] . Um nun herauszufinden, was im Text steht, wird entweder auf bereits vorhandene Transkriptionsdaten zurückgegriffen oder eigens eine Transkription angefertigt. Diese Vorgehensweise der Texterkennung von Handschriften hat sich grösstenteils durchgesetzt. Nun wird vor allem versucht, den Leistungsumfang der HTR (Handwritten Text Recognition) zu erhöhen. Der Grundsatz, dass die grösstmögliche Menge von Trainingsdaten die Fehlerquote am stärksten senkt, setzt viel Vorarbeit des Forschenden voraus. Weil jede Handschrift anders ist, kann man keine Mindestseitenzahl angeben, die transkribiert sein muss, um ein zuverlässiges HTR-Modell zu erhalten. Aus diesem Grund dienen Schriften oder Sprachen, die der Software bereits früher antrainiert wurden, als Grundlage für eine effizientere Transkription, in dem mit sogenannten Basismodellen operiert wird [9] .

2. Praktischer Teil: Umsetzung und Stolpersteine

Mit diesem Vorwissen ging es nun darum, einen zu bearbeitenden Bestand ausfindig zu machen. Nicht nur mussten datenschutzrechtliche Faktoren berücksichtigt werden, es musste auch ein Bestand mit einer durchgehend gleichförmigen Struktur eruiert werden. Folgende Kartei der früheren Staatsarchivarin des Kantons Thurgau, Verena Jacobi (Amtszeit 1979-1986) entsprach beiden Grundvoraussetzungen.

Die Kartei «Grossratsmitglieder (Kartei Jacobi) 1803 bis ca. 1930» mit der Signatur StATG 2'203'1 verzeichnet annähernd sämtliche Mitglieder des Grossen Rats seit der Kantonsgründung im Jahr 1803 bis circa ins Jahr 1930. Eine Karteikarte entspricht in der Regel einem Ratsmitglied. Bei den frühsten Amtsperioden zwischen 1803 bis 1869 kommt es vor, dass aufgrund fehlender oder gleicher Vornamen im Grossratsprotokoll auf einer Karte zwei Personen notiert sind, weil man sie nicht klar unterscheiden kann. Neben dem vollen Namen sind auf der Karteikarte die Amtszeiten, die zusätzlichen Ämter, die Herkunft oder der Wohnort und eventuell der Beruf erwähnt.

Abb. 1: Beispiel einer Karteikarte mit handschriftlichen Notizen.

2.1 Struktur

Nach der Festlegung auf den Bestand wurden die 1260 Karteikarten untersucht, um den Aufbau und den Informationsgehalt einzuschätzen. Die Karteikarten sind einseitig beschriftet, hauptsächlich mit Schreibmaschine, teilweise aber auch

handschriftlich mit Bleistift. Die Informationen zur Ratstätigkeit der jeweiligen Person sind durchgehend mit Schreibmaschine geschrieben worden. Zusätzliche Informationen, die eher anekdotischen oder verweisenden Charakter haben, wurden von Verena Jacobi von Hand hinzugefügt. Besonders die Karteikarten von Personen, die sie nicht eindeutig identifizieren konnte, sind oft mit einer handschriftlichen Anmerkung oder einem Verweis versehen.

Die Struktur der Karteikarten ist sehr flach, es gibt keine Hierarchie innerhalb der Kartei. Sie sind alphabetisch nach Nachnamen geordnet, sämtliche Informationen müssen via den Einstieg über die Person ausfindig gemacht werden. Weil die Informationen an die Person geknüpft sind, müsste beispielsweise bei der Suche nach der Zusammensetzung des Grossen Rats in einem bestimmten Jahr oder nach besonderen Ämtern und Aufgaben der gesamte Bestand durchgesehen werden. Die Angaben befinden sich zudem nicht immer an exakt derselben Stelle auf der Karteikarte, da sie nicht am vorgesehenen Ort Platz hatten oder möglicherweise später hinzugefügt wurden.

2.2 Scanvorgang

Da auf den Karteikarten neben dem schwarz der Schreibmaschinentinte auch hellgraue Bleistiftanmerkungen zu sehen sind, konnte die Vorlagenart «schwarz- weiss» nicht verwendet werden, obwohl dies kleinere Dateivolumen ergeben hätte. Die Bleistifteinträge sind bei diesem hohen Kontrast im Scan für das menschliche Auge nur unzureichend sichtbar und werden auch von Transkribus nicht mehr zufriedenstellend erkannt. Deshalb wurde auf die Vorlagenart «Graustufen» ausgewichen, die grössere Dateien produziert, aber Schattierungen besser abbildet. Trotzdem ist es hilfreich, die Scans möglichst kontrastreich zu halten, um Verunreinigungen und Staub auszublenden. Wie oben beschrieben, können dann aber neben den störenden Punkten auch bedeutungstragende «Flecken» wie Bleistiftstriche oder farbliche Hervorhebungen verloren gehen.

2.3 Auswahl der notwendigen Informationen

Bei der Auswahl der notwendigen Informationen für die Titelvergabe ergab die Durchsicht der Benutzeranfragen zu den Grossratsmitgliedern folgende Voraussetzungen: der Dossiertitel muss den Vor- und Nachnamen enthalten, etwas über die Herkunft der Person aussagen und die Zeitspanne angeben, in dem die Person im Grossen Rat gesessen ist.

Ebenso soll der Titel aus einem grammatikalisch korrekten Satz bestehen, also keine Aufzählung von Stichwörtern sein. Mit diesem Aufbau können sowohl das Interesse an einer Einzelperson, an der geographischen

Verteilung der Legislativpolitiker als auch an der Zusammensetzung des Grossen Rats zu einem bestimmten Amtsjahr bedient werden.

Obwohl die Übernahme sämtlicher Informationen auf den Karteikarten den Dossiertitel nicht sprengen würde, soll dieser Bestand als Beispiel für zukünftige Bestände dienen, die mehr Informationen enthalten, als für die Titelvergabe benötigt werden. Neben der Überprüfung der Genauigkeit der Transkription soll am Ende auch eine Aussage darüber getroffen werden, ob die Software die gewünschten Informationen gefunden und ausgewählt hat.

2.4 Segmentierung

Die Segmentierung eines Scans bewirkt, dass die Texteinheit einer Seite in kleinere Textblöcke und Textzeilen unterteilt wird. Statt dass die Texterkennung wie bei der OCR-Methode jedes Zeichen als einzelnes Symbol losgelöst von seinen jeweiligen Nachbarn links und rechts erkennt, wird eine «Grundlinie» definiert. Diese Grundlinie besteht aus mehreren Zeichen und entspricht einer ganzen Zeile oder einem Teil davon. Die einzelnen Pixel werden bei der Definition der Grundlinie nur einer Zeile, nicht aber bereits einem Buchstaben zugeordnet. Nach dem Abschluss der Segmentierung ist somit noch unklar, welche Symbole auf(?) einer Grundlinie vorkommen, und welchen Inhalt die Zeile vorweist. Dabei ist zu beachten, dass eine nicht erkannte Textzeile nicht transkribiert werden kann.

Die Beschränkung auf die Bestimmung der Grundlinie im ersten Schritt ermöglicht es, das Layout des Textes zu erkennen. Besonders bei formularhaften Dokumenten und streng strukturierten Texten helfen die Grundlinien, den Aufbau der gesamten Serie nachzuvollziehen, bevor der Inhalt der einzelnen Fälle miteinbezogen wird. [10]

Die Grundlinien der Grossratskartei waren zum grössten Teil korrekt identifiziert. Die gesperrte Schreibweise der Nachnamen bereitete teilweise Probleme, weil die Namen nicht als Einheit erkannt wurden und die Grundlinien dann manuell miteinander verbunden werden mussten.

Nachdem die Grundlinien von Transkribus zu einem grossen Teil korrekt identifiziert worden waren, galt es, diese einzelnen Linien «Textregionen» zuzuordnen. Diese Textregionen dienen dazu, die Grundlinien in strukturelle oder inhaltliche Gruppen einzuteilen. So werden beispielsweise alle Übertitel einer Textregion zugewiesen oder Personennamen auf den Scans eingegrenzt. Weil bestimmte Teile der Karteikarten später für die automatische Titelvergabe verwendet werden sollen, entsprachen die Textregionen den inhaltlichen Titelvorgaben. Um trotzdem sämtliche Informationen auf der Karteikarte klassifizieren zu können, kamen

inhaltliche Kategorien dazu, die später keine Verwendung im Titel finden werden. Folgende sechs Textregionen haben sich dabei ergeben:

Name, Wohnort, Dauer der Mitgliedschaft im Grossen Rat, Beruf, Verweis auf Quellen und andere Personen und Diverses. Die Textregion Diverses wurde geschaffen, um jene Informationen, die keiner der übrigen fünf Textregionen zugewiesen werden konnten, aufzufangen. Sie beinhaltet so verschiedene Angaben wie Ämter ausserhalb des Grossen Rats, Gründe für das Ausscheiden aus dem Grossen Rat oder die Parteizugehörigkeit. Weil die beiden letzteren Regionen nur vereinzelt auf den Karteikarten vorkommen, wurden sie nicht weiter berücksichtigt.

Die Textregionen, welche später für die Titelgenerierung wichtig sein werden, sind wie oben bereits notiert, die ersten drei. Anhand dieser kann die Person mit einigen Vorbehalten identifiziert werden.

Weil die Karteikarten sehr stark strukturiert, gleichzeitig die Informationen aber nicht konsequent am selben Ort auf der Karteikarte vorhanden sind, war es zu erwarten, dass ein gutes Modell für die automatische Lokalisierung von Textregionen viele von Hand bearbeitete Karteikarten erfordern würde. Um den Versuch weder zeitlich noch personell zu überstrapazieren und ihn gleichzeitig aussagekräftig zu gestalten, musste eine Obergrenze für die Anzahl der händisch bearbeiteten Scans gesetzt werden. Jeweils 250 Karteikarten, also circa ein Fünftel des Bestands, sollten in diesem Schritt und in allen weiteren höchstens bearbeitet werden, bis ein Modell trainiert wird. Transkribus ist im Hinblick auf die Erkennung von linearen Texten entwickelt worden, die weniger Strukturen im Aufbau aufweisen. Fliesstexte sind, auch wenn sie Überschriften, Randnotizen und Fussnoten enthalten, ohne Abbildung des Layouts grösstenteils verständlich. Im Gegensatz dazu sind die Karteikarten und besonders der Versuch der automatischen Titelgenerierung vom Layout abhängig, weil die Position des Textes etwas über dessen Inhalt aussagt.

Abb. 2: Die von Hand gezogenen rechteckigen Textregionen müssen jeweils einer der sechs gewählten Kategorien zugewiesen werden. Im Reiter Layout kann die Zuordnung überprüft werden.

2.5 Transkription

Viele Karteikarten sind ausschliesslich mit Schreibmaschine verfasst worden. Deshalb könnte man davon ausgehen, dass für den vorliegenden Bestand ein schon vorhandenes Transkriptionsmodell für Schreibmaschinenschriften verwendet werden könnte. Weil aber gerade die spärlichen, oft nur schwer lesbaren, handschriftlichen Bleistiftnotizen genauere Auskunft darüber geben, ob eine Karteikarte doppelt vorkommt oder teilweise Informationen zu den drei ausgewählten Pflichttextregionen enthält, musste ein eigenes Modell trainiert werden. Nachdem die Grundlinien zufriedenstellend erkannt und gegebenenfalls korrigiert wurden, folgt nun dieser nächste Schritt.

Im Interface folgt der Benutzer den Grundlinien und versucht, so genau wie möglich das Markierte zu transkribieren. [11]

Abb. 3: Je eine Grundlinie muss transkribiert werden. Die gerade bearbeitete Linie wird im Scan und in der Transkription blau unterstrichen.

2.6 Modelltraining

Nachdem ein Fünftel der Karteikarten segmentiert und transkribiert war, ging es nun zum Modelltraining über. Die Modelle sind für das Gelingen der automatisierten Dossiertitelvergabe essenziell. Sind sie genau genug, ist der noch unbearbeitete Teil der Karteikarten innert kürzester Zeit bereit für den Export aus Transkribus.

Um Aussagen über den Arbeitsaufwand für das Modelltraining zu machen, wurden drei unterschiedlich grosse transkribierte Textkorpora ausgewählt und jeweils ein Modell damit trainiert. Die transkribierten Scans sind in zwei Gruppen eingeteilt: das «Train Set», das Trainingsdatenset, und das «Validation Set», dessen Genauigkeit

im Abgleich mit dem Trainingsdatenset bestimmt wird. Die Lernkurve stellt graphisch dar, wie sich die Character Error Rate (CER) mit ansteigendem Datenvolumen verbessert. Liegt beim ersten Modell die CER noch bei 9,49 %, ist das dritte Modell mit einem viermal so grossen Grundstock an Trainingsdaten bereits unter 5 %. Viel tiefer als die 4,88 % wird die CER bei diesem Bestand vermutlich nicht sinken, weil die wenigen handschriftlichen Einträge die sehr hohe Genauigkeit der Maschinenschrifterkennung stören.

Abb. 4: Das dritte Modell besteht aus 194 Karten als Trainingsdatenset und aus 24 Karten zur Überprüfung der Transkriptionsgenauigkeit und ergibt eine CER von 4.88%.

2.7 Export der Transkription

Die Exportmöglichkeiten in Transkribus sind sehr vielfältig. Grundsätzlich kann man zwei Arten unterscheiden, wie die Transkriptionen aus Transkribus exportiert werden. Einerseits wird nur die Transkription exportiert, um wie in diesem Fall den Text weiter zu bearbeiten. Andererseits ist es auch möglich, die Transkription verknüpft mit dem Scan zu exportieren, um das Original direkt mit der

Transkription vergleichen zu können. Weil für dieses Vorgehen nur der Inhalt einiger Textregionen von Interesse ist und die Lokalisation auf dem Scan der Karteikarte für die Titelvergabe keine Rolle spielt, konnte ein Exportformat ausgewählt werden, das rein textbasiert ist. Da xml-Dateien mithilfe eines Editors nachbearbeitet werden können, fiel die Entscheidung auf das PageXML Format.

Abb. 5: Die Exportmöglichkeiten aus Transkribus sind vielfältig und können sehr genau auf die eigenen Bedürfnisse abgestimmt werden.

2.8 Auswahl der benötigten Angaben aus xml

Die entstandenen xml-Dateien pro Seite und Karteikarte mussten nun in einem xml- Editor geöffnet werden, um dann mit einem Skript die benötigten Informationen herauszufiltern. Der Editor ermöglicht es, die benötigten Angaben aus der xml- Datei mithilfe einer XSL-Transformation zu extrahieren und als einfachen Text oder als eine Excel-Tabelle abzuspeichern.

Der Output bestimmt, dass nur Text herausgefiltert wird und beispielsweise die Koordinaten der einzelnen Grundlinien ausgelassen werden sollen. Innerhalb des Texts sollen nur die Textregionen ausgewählt werden, die einen bestimmten Tag besitzen.

Dieser Vorgang wird im Oxygen-Editor mit einem Befehl für sämtliche xml- Dateien in einem Ordner durchgeführt. Danach konnte der gewünschte Text mithilfe des Skripts herausgefiltert werden.

Abb. 6: Die exportierte Transkription aus Transkribus auf der linken Seite und das Skript für die Auswahl der benötigten Textregionen auf der rechten Seite.

Der herausgefilterte Text wurde als .txt-Datei abgespeichert, mit Notepad ++ wurden die überzähligen Zeilenumbrüche entfernt. Zusätzlich filtert man in Excel die regulären Ausdrücke, um eine standardisierte Reihenfolge herzustellen. Erste Aussagen über den endgültigen Titel liessen sich nun machen: die daraus resultierende Liste entsprach bei weitem nicht den Qualitätsstandards der manuellen Titelvergabe. Wenn eine Textregion falsch, nicht oder doppelt erkannt wird, dann verschieben sich die gewünschten Felder. Anstatt der Ordnung Name, Ort und der Amtszeit im Grossen Rat wurde dann beispielsweise der Name übersprungen und es erschien als dritte ausgewählte Textregion die Kategorie Beruf.

Die Nachbearbeitung des exportierten Texts in Excel kann aufgrund von Fehlern in den Transkriptionen und in der Zuweisung der Textregionen unbestimmt ausufern. Umso wichtiger ist es darum, in die ersten Arbeitsschritte genügend Zeit zu investieren, um später nicht allzu viele Einzeleinträge korrigieren zu müssen.

2.9 Import in Archivinformationssystem

Der Import der Titel in das Archivinformationssystem ScopeArchive wurde über den Scope Findmittel-Assistenten bewerkstelligt. Das Zusatzprogramm wandelt .txt- oder xml-Dateien so um, dass damit Felder in ScopeArchive befüllt werden können. Nachdem die Felder der Dossiermaske befüllt worden sind, müssen nun die Hierarchiestufen zugeordnet werden. Die Satz- und Sonderzeichen dienen dabei als

Trenner und ermöglichen es dem Assistenten, aus der Textdatei eine hierarchische Struktur zu schaffen. Dies sieht folgendermassen aus:

Abb. 7: Die importierten Dossiertitel in ScopeArchive.

Das Endresultat kann so leider nicht aufgeschaltet werden. Die Titel enthalten noch zu viele Fehler. Ebenso sind beim Export der

Transkription und der darauffolgenden Bearbeitung in verschiedenen Editoren Buchstaben und Ziffern verloren gegangen, die für das Verständnis notwendig sind. Es wäre vertretbar, wenn gewisse Titel mehr Informationen enthalten würden als andere, weil beispielsweise der Beruf als Ort erkannt worden ist und deshalb auch noch aufgeführt wird. Solange aber die Titel so heterogen ausfallen und teilweise nicht lesbar sind, weil die kombinierten Transkriptions-, Segmentations- und Exportfehler zusammen einem Buchstaben- und Ziffernsalat gleichen, muss an der Qualität der Rohdaten gearbeitet werden. Dies bedeutet, dass den Transkriptions- und Segmentationsmodellen eine grössere Basis an Scans zur Verfügung gestellt werden muss, um Fehler zu vermeiden, die sich mit jedem weiteren Schritt Richtung automatisierte Titelvergabe multiplizieren.

Trotzdem ist es bemerkenswert, dass gewisse Dossiertitel praktisch fehlerfrei sind und ins Archivinformationssystem übernommen werden könnten.

3. Abwägungen betreffend den finanziellen und personellen Aufwand

Die Vorbereitungsarbeiten vor der Erschliessung des vorliegenden Bestandes haben einige Zeit in Anspruch genommen, die bei der Abwägung des Aufwands miteinbezogen werden müssen. Die Erfahrung, die für die speditive Abwicklung von manuellen Erschliessungsprojekten durch die tägliche Arbeit bereits vorhanden ist, fehlte beim Versuch mit den beschriebenen Programmen komplett.

Bei der Durchsicht der Karteikarten stellte sich heraus, dass sie gar nicht so streng strukturiert aufgebaut waren, wie sie im ersten Augenblick den Anschein gemacht hatten. Für diesen Versuch war dies in Anbetracht der restlichen Erschliessungsvorgaben (ein Fünftel des Bestands muss als Trainingsgrundlage für die Modelle dienen) nicht umsetzbar.

Beim vorliegenden Versuch sind mehrere Voraussetzungen nötig, um das Gelingen zu garantieren. Zuerst muss man sich mit dem HTR-Programm Transkribus vertraut machen. Es besitzt eine anwenderfreundliche Oberfläche, ähnlich wie man es von den Office-Produkten von Microsoft kennt.

Der Vorgang der Transkription und der Identifizierung der Textregionen kann von jeder Person bewerkstelligt werden, die Handschriften lesen kann. Danach ist, wie auch bei der Erschliessung von Hand, grosse Ausdauer und Aufmerksamkeit gefragt, um eine möglichst fehlerfreie Transkription anzufertigen. Nachdem die Grundlinien, die Transkription und die Textregionen kontrolliert worden sind, geht es darum, mit Texteditoren umgehen zu lernen und ihre Fähigkeiten auszuloten.

Die Kombination der je nach Struktur des Bestands umfassenden Bearbeitungszeit inklusive Controlling und dementsprechenden Lohnkosten und der Abrechnung pro Seite von Transkribus setzt nicht unbedingt Anreize, um diese Software für ein archivisches Projekt zu

benutzen. Trotzdem muss es im langfristigen Interesse des Archivs liegen, bei der Erschliessungspraxis den grossen Kostenfaktor Lohnkosten zu überdenken und gegebenenfalls effizienter einzusetzen.

4. Controlling: Qualitätsansprüche an die Titelvergabe

Vorgaben wie die Ansprüche der Archive an den Informationsgehalt des Titels, die Schreibweise von Namen und die gleichförmige Verwendung von Abkürzungen und Begriffen über einen langen Zeitraum hinweg müssen je nach Bestand und Datenlage aktualisiert werden. Oft sind dabei Kompromisse nötig, um den Erschliessungsaufwand in ein realistisches Verhältnis mit der Aussagekraft des Bestandes zu bringen. Lange, zu detailreiche Titel, die schwer lesbar sind, oder sehr viele Felder, die nicht konsequent befüllt werden, verlangsamen die Erschliessungsarbeit und schmälern die Qualität des Endergebnisses. Weil die Dossiertitelvergabe bei der computergestützten Erschliessung auf möglichst

gleichförmige Titel angewiesen ist, müssen im direkten Vergleich mit der manuellen Titelvergabe gewisse Abstriche bei der Qualität der Titel in Kauf genommen werden. Es wäre denkbar, dem Dossier im Titel eine Sammlung von Stichwörtern anzugliedern, die, statt einen starren Satz zu bilden, den Inhalt näher beschreiben. Ähnlich wie dies in Bibliothekskatalogen gehandhabt wird, könnte man, angelehnt an die verbale oder klassifikatorische Sacherschliessung, die Dossiers verschlagworten. Die Verschlagwortung bietet, im Gegensatz zu den Dossiertiteln, den Mehrwert der

Verknüpfung der Schlagwörter untereinander. [12]

5. Schlussbemerkungen

Der Versuch ist geglückt: der analogen Karteikarte aus den Jahren um 1970 wurde ein vom Computer generierter Dossiertitel zugewiesen, ohne dass sich die Karte jemals in Archivarenhänden befunden hätte, vom Scanprozess abgesehen. Doch der Titel genügte den Qualitätsansprüchen, die man sich von der Erschliessung von Hand gewohnt ist, auf grosser Strecke nicht. Zu klein war die Menge an Trainingsdaten, um zuverlässige Modelle für die automatisierte Segmentierung und Transkription zu erstellen, zu heterogen der Aufbau der Karteikarten an sich, um die benötigten Informationen fehlerfrei zu lokalisieren. Genau hier befindet sich einer der Ansatzpunkte, der mit der stetigen Digitalisierung der ABD-Welt adressiert werden muss. Statt an der herkömmlichen Art der Titelvergabe festzuhalten, lohnt es sich, mit der vorhandenen Hard- und Software zu experimentieren und mit den Datenschutzgesetzen im Blick zu schauen, was momentan machbar ist. Ältere gleichförmige Grossserien wie handgeschriebene Regierungsratsbeschlüsse aus dem

19. Jahrhundert liessen sich so beispielsweise erschliessen.

Die Aufteilung der Karteikarten in Textregionen hat sich als zu ungenau erwiesen. Dem Modell stand 1/5 des Bestands als Grundlage zur Verfügung, um den Aufbau der Karteikarten zu trainieren. Diese Anzahl von manuell eingegebenen Textregionen reichte aber nicht aus, um zuverlässig Informationen auf den unbearbeiteten Karteikarten zu identifizieren. Einerseits hätte man dieses Problem beheben können, indem man die Trainingsgrundlage für das Modell bedeutend vergrössert hätte. Dies war aber unter der Vorgabe, nur eine begrenzte Anzahl der Karteikarten für die einzelnen Arbeitsschritte aufzubereiten, nicht sinnvoll. Der Bestand ist offensichtlich zu unregelmässig, um mit dem vorhandenen Trainingsmaterial zufriedenstellende Ergebnisse zu liefern.

Wie stark würde sich beispielsweise der Umbruch im Archiv manifestieren, wenn sich im Dossiertitel statt eines grammatikalisch korrekten Satzes neu eine Stichwortauflistung befände? Die Suchgewohnheiten der Menschen haben sich im

Zeitalter von Google und Co. so stark verändert, dass wir problemlos mit Worthaufen umgehen können. Lässt sich die Vergabe von Stichworten zusätzlich an einen Computer auslagern, kann die Erschliessungsarbeit neu gedacht werden. Massenakten könnten erschlossen werden, *weil* sie einen grossen Umfang haben und gleichförmig aufgebaut sind und müssten nicht zuwarten, bis Sondererschliessungsprojekte genehmigt werden.

Um diesem Ziel näher zu kommen, müssen in der Zwischenzeit Erfahrungswerte gesammelt und einige Anfangsschwierigkeiten ausgebügelt werden. Einerseits ist es nötig, die Bestände sorgfältig auszuwählen und die Modelle genauer zu kontrollieren, um aufwändige Nachbearbeitungen zu verhindern. Andererseits wäre die Zusammenarbeit mit einer in der Informatik versierten Person hilfreich, um bei der Bearbeitung der Unterlagen möglichst wenig Datenverlust einzufahren.

Literaturverzeichnis

Bonhomme, Marie-Laurence: Répertoire des Notaires parisiens. Segmentation automatique et reconnaissance d'écriture. Rapport exploratoire. Paris 2018.

Gantert, Klaus, Hacker, Rupert: Bibliothekarisches Grundwissen. München 2008.

Mühlberger, Günter: «Archiv 4.0 oder warum die automatisierte Texterkennung alles verändern wird.» In: Deecke, Klara, Grothe Ewald (Red.): Massenakten – Massendaten. Rationalisierung und Automatisierung im Archiv. 87. Deutscher Archivtag in Wolfsburg. Tagungsdokumentationen zum Deutschen Archivtag 22, (2018), 145-156.

170.7 Gesetz über den Datenschutz (TG DSG) vom 09.11.1987, Stand 01.07.2012. http://www.rechtsbuch.tg.ch/frontend/versions/386. Zugriff am 14.07.2020. https://www.abbyy.com/de/finereader/what-is-ocr/. Zugriff am 05.09.2020.

Fußnote

1 Bonhomme, Répertoires des notaires parisiens, S. 3.
2 Mühlberger, Archiv 4.0, S. 145f.
3 Bonhomme, Répertoires des notaires parisiens, S. 3.
4 Gesetz über den Datenschutz (TG DSG), § 6.
5 https://www.abbyy.com/de/finereader/what-is-ocr/. Zugriff am 05.09.2020.
6 Bonhomme, Répertoire des Notariats parisiens, S. 7.
7 Mühlberger, Archiv 4.0, S. 145.
8 Bonhomme, Répertoire des Notaires parisiens, S. 6.
9 Mühlberger, Archiv 4.0, S. 146.
10 Mühlberger, Archiv 4.0, S. 149.
11 Mühlberger, Archiv 4.0, S. 150.
12 Gantert, Bibliothekarisches Grundwissen, S. 177-179. Quellenverzeichnis StATG 2'203'1, Grossratsmitglieder (Kartei Jacobi), 1803 bis ca. 1930.
Quellenverzeichnis
StATG 2'203'1, Grossratsmitglieder (Kartei Jacobi), 1803 bis ca. 1930.