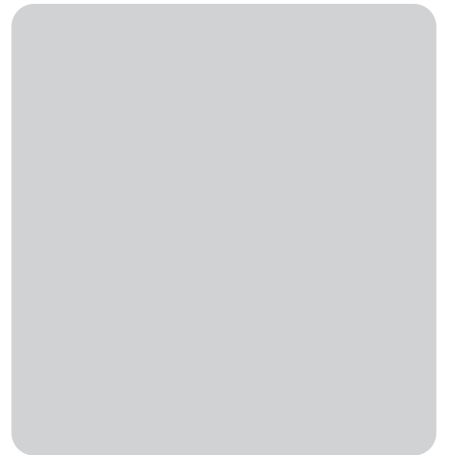


White Paper



ABBYY® Historic OCR

Grundlagen und Herausforderungen bei der Verarbeitung historischer Dokumente mit Fraktur-OCR

Michael Fuchs, ABBYY Europe (München)

Historische Dokumente in einer „digitalen“ Welt

Für den modernen Menschen gehören Computer, Internet und die elektronische Speicherung von Information längst zum Alltag. Als klassische Bewahrer gedruckter Dokumente können sich Bibliotheken dem Strudel der weltweiten Digitalisierung heute nicht mehr entziehen. Daher wurde schon vor einigen Jahren mit dem Aufbau digitaler Bibliotheken begonnen, z.B. die Deutsche Digitale Bibliothek oder die europeana. Dies ist die Antwort auf das Bedürfnis, auf Informationen von überallher und jederzeit zugreifen zu können.

Die Prozesse, mit denen sich bereits digital entstandene Dokumente („digital born files“) für den elektronischen Zugriff und die Recherche per Internet bereitstellen lassen, unterscheiden sich grundlegend von der Aufbewahrung, dem Durchsuchen und dem Aufrufen von auf Papier gedruckten Texten. Um auch diese papierbasierten Informationen in unser „digitales Leben“ zu integrieren, hat sich in den letzten Jahren „Optical Character Recognition“ (OCR) als wichtige Technologie etabliert. Moderne gedruckte Dokumente lassen sich mit der heute zur Verfügung stehenden hoch automatisierten OCR-Software sehr viel komfortabler, schneller, kostengünstiger und auch zuverlässiger digitalisieren als wenn sie von Hand abgetippt würden.

Universitäten, Bibliotheken oder Archive, die ihre historischen Dokumentenbestände digitalisieren wollen, stehen immer noch vor einer großen Herausforderung, denn die Erfassung und die zuverlässige Erkennung von historischen Dokumenten ist nicht trivial: Die über viele Jahrhunderte verwendeten „gebrochenen Schriften“ wie „Textur“, „Rotunda“ oder „Fraktur“, die einen wichtigen Bestandteil der europäischen Buchdrucktradition darstellen und damit wesentlich für unser europäisches Kulturerbe sind, widersetzen sich häufig einer automatisierten Erkennung. Denn neben schlechten Papierqualitäten der historischen Vorlagen existierten beim Druck älterer Druckdokumente keine standardisierten Schriften. Daher weicht oft das Aussehen gleicher Buchstaben von Dokument zu Dokument stark voneinander ab. Auch Sprache und Rechtschreibung haben sich in den letzten Jahrhunderten stark verändert, dadurch können aktuelle Wörterbücher oft nicht zu einer automatischen Korrektur herangezogen werden.

Die folgenden Kapitel beleuchten die Hintergründe, die bei der Erfassung historischer Dokumente durch moderne OCR-Technologien eine Rolle spielen.

Wie optische Zeichenerkennung funktioniert

... und warum alte Dokumente eine große Herausforderung darstellen

„Optical Character Recognition“ heißt auf Deutsch „optische Zeichenerkennung“. Dabei handelt es sich bei OCR um eine höchst komplexe Abfolge mathematischer und linguistischer Prozesse. Die wichtigsten Schritte, die für eine optimale Digitalisierung (historischer) Dokumente notwendig sind, gliedern sich wie folgt:

Schritt 1 – Von der Bilderfassung bis hin zur Bildoptimierung

Die erste – und auch eine der wichtigsten – Voraussetzungen für eine zuverlässige Erfassung historischer Dokumente durch OCR-Technologie ist die Erstellung eines Scans in bestmöglicher Qualität. Die digitalen Pixelbilder können vom Originaldokument (einzelne Blätter oder ein Buch) oder auch von Mikrofilmen erstellt werden. In der Regel werden dazu spezielle Scanner verwendet, die auch in der Lage sind „schwieriges, analoges Material“ in hoher Qualität zu digitalisieren.

Damit für den OCR-Vorgang genügend „Rohdaten“ zur Verfügung stehen sollten/müssen bestimmte Mindestanforderungen erfüllt werden: Eine Dokumentenseite mit normalen Schriftgrößen muss mit einer Auflösung von mindestens 300 dpi und vorzugsweise in Graustufen oder in Farbe vorliegen. Durch ein einfaches „bi-tonales“ Scannen in Schwarz-Weiß könnten wichtige

Informationen der Vorlage entfallen, die dann der OCR-Engine nicht mehr zur optimalen Entschlüsselung des Textes zur Verfügung stehen. An einem einfachen „Bitmap“-Scan in Schwarz-Weiß lassen sich z. B. viele der nachfolgend beschriebenen Bildoptimierungsmöglichkeiten gar nicht mehr durchführen.

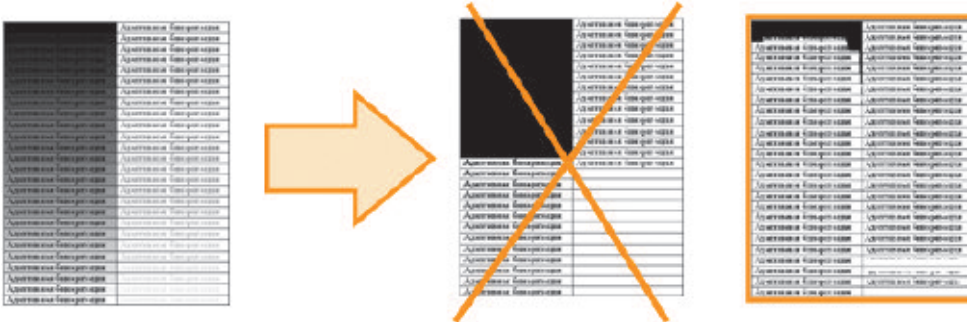
Denn zur optimalen Vorbereitung für die Erfassung durch OCR müssen die Dokumente, nachdem sie digitalisiert und in Bilddateien elektronisch abgelegt sind, zunächst noch einer weiteren Vorverarbeitung unterzogen werden: Dazu gehört z. B. bei Bildern mit zu geringer Auflösung die Skalierung und Anpassung ihrer Auflösung auf mindestens 300 dpi, die Trennung von Doppelseiten, die Rotation um 90, 180 oder 270 Grad (damit alle Dateien in gleicher Ausrichtung vorliegen) und die Bildbeschneidung („Cropping“). Während der Bildvorverarbeitung werden auch typische Scan-Fehler so weit wie möglich korrigiert, beispielsweise über das automatische Geraderücken von Dokumentseiten („De-Skewing“), durch die automatische Begrädigung von Textzeilen sowie eine kontrollierte automatische Entfernung von Staub sowie Hintergrundrauschen. Hierbei muss allerdings genau darauf geachtet werden, dass diese nachträglichen Eingriffe nicht zu stark ausfallen, damit im Zuge der Bildoptimierung nicht gleichzeitig auch kleinste Zeichen, wie z. B. Punkte bei i, ä, ö oder ü mit entfernt werden.

Die eigentliche OCR findet immer auf der Ebene eines sogenannten „Binärbildes“, einer schwarz-weißen Rastergrafik, statt. Um diese zu erzeugen, nutzt die OCR-Technologie von ABBYY eine intelligente Hintergrundfilterung mit adaptiver Binarisierung. Dabei wird das Bild mit verschiedenen Algorithmen analysiert und dann schrittweise in ein Schwarz-Weiß-Bild überführt, sodass der Text vom Hintergrund getrennt wird und die einzelnen Buchstaben vollständig und nicht zu fett erhalten bleiben.

Wie in Abbildung deutlich wird, können bei der Digitalisierung gerade von Dokumenten mit wenig Kontrast oder ungleichmäßigen Texthintergründen durch eine falsche Binarisierung wertvolle Textinformationen verloren gehen. Denn der Schwellenwert der Binarisierung hat einen direkten Einfluss auf die Qualität der einzelnen Buchstaben und wirkt sich damit unmittelbar auf die Gesamtqualität des OCR-Ergebnisses aus.



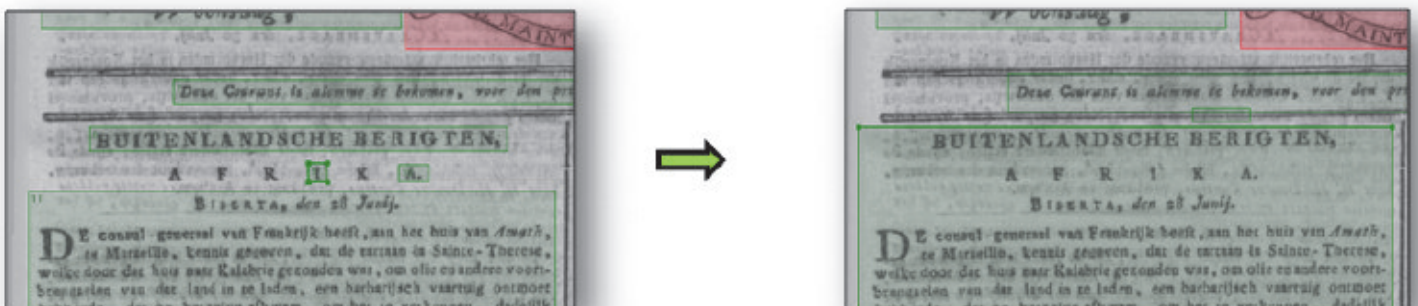
Durch eine falsche Binarisierung können für die OCR-Engine wertvolle Textinformationen während der Digitalisierung verloren gehen.



Schritt 2 – Die Dokumentenanalyse

Für die Layoutanalyse des jetzt als Bilddatei vorliegenden (historischen) Dokumentes kommt viel Mathematik zum Einsatz. Moderne Dokumente, sind meist sehr klar und übersichtlich gegliedert und bereiten daher oft auch bei der automatischen Layouterkennung keine Probleme. Historische Dokumente und Zeitungen hingegen wurden meist „kreativer“ gestaltet und haben oft keinerlei standardisierte Layouts. Die Identifikation der einzelnen Elemente in Vorlagen stellt daher hohe Anforderung an die OCR-Technologie:

Sie muss in diesem Schritt genau erfassen, aus welchen Komponenten sich die vorliegende Bilddatei zusammensetzt – die OCR-Engine definiert dabei ganz exakt, aus welchen Bereichen – Bilder, Text, Tabellen oder ähnliches – sich das Layout zusammensetzt. Deshalb ist es wichtig, zur Entschlüsselung historischer Dokumente eine entsprechend optimierte Analyse-Technologie einzusetzen. Nur wenn die Textbereiche korrekt und vollständig definiert werden, kann das Endergebnis des OCR-Prozesses zufriedenstellend sein.



Werden die Textblöcke eines Layouts nicht korrekt erkannt, fehlen diese Informationen auch nach der Texterkennung. Die Abbildung zeigt Layoutanalyse-Verbesserungen auf alten Zeitungen, die ABBYY während des IMPACT-Projektes erzielt hat.

Schritt 3 – Die Zeichenerkennung

Nach Definition der unterschiedlichen Erkennungsbereiche startet die eigentliche Zeichenanalyse, die sich von der Identifikation der einzelnen Zeilen bis hin zur Auffindung der einzelnen Buchstaben und Zeichen vortastet. Das „Auslesen“ basiert dabei auf sogenannten „Classifiern“. ABBYY OCR-Technologie verwendet dazu unterschiedlichste Algorithmen, um die Buchstaben möglichst korrekt zu erkennen und die ausgelesenen Ergebnisse – zunächst noch Hypothesen – zur Verifizierung mit im System hinterlegten Muster-Zeichen zu vergleichen:

- „Raster Classifier“ etwa vergleichen die Pixel der einzelnen Buchstaben mit den hinterlegten „Master“-Zeichen,
- „Kontur Classifier“ werten die Buchstaben-„Umrisse“ nach vorhandenen Mustern aus.
- „Struktur Classifier“ reduzieren die Buchstaben auf ein einfaches Vektor-Skelett, das ebenfalls auf Ähnlichkeit mit hinterlegten „Master“-Zeichen überprüft wird.
- „Spezialisten Classifier“ werden genutzt, um ähnliche Zeichen klar voneinander unterscheiden zu können z.B. „B“ von „D“ oder „D“ von „O“

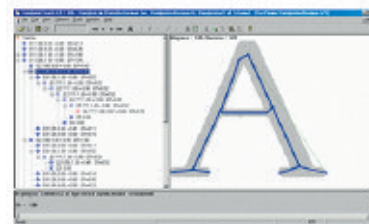
Raster classifier



Contour classifier



Structure classifier



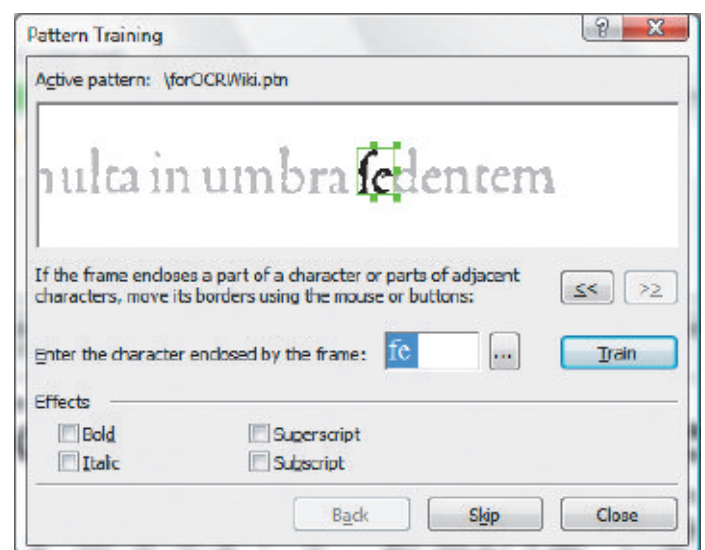
Feature differentiating classifier



ABBYY verwendet eine Vielzahl spezieller „Experten“ um einzelne Zeichen zu erkennen und so optimale Erkennungsergebnisse zu erzielen.

Erkennen individueller Sonderzeichen

Für das Erkennen von Standardzeichen nutzt OCR-Technologie üblicherweise typische „Patterns“ (Muster), die bereits in der Software hinterlegt sind (siehe oben). Darüber hinaus können sich viele OCR-Lösungen auch individuelle, für ein Dokument oder einen Dokumententyp spezifische Sonderzeichen, die im softwareeigenen Mustervorrat noch nicht vorhanden sind, während der Analyseprozesse nach und nach „antrainieren“. Dies erzielt vor allem bei der Erkennung untypischer Zeichen oder auch spezieller Schmucksymbole sehr gute Erfolge. Solche „individuellen Muster“ werden immer in Kombination mit bereits hinterlegten Classifiern verwendet und steigern durch die damit zusätzlich zur Verfügung stehenden Analyseinformationen die Erkennungsrate auch bei schwierigen historischen Textvorlagen. Technisch muss angemerkt werden, dass für das Trainieren von Zeichen die Bildauflösung der verarbeiteten Bilder und der Mustervorlagen übereinstimmen muss, da dieser Optimierungsansatz nur funktioniert, wenn die Zeichenhöhe in Pixel übereinstimmt. Da „trainierte“ Zeichen nur ein weiteres Bewertungsmerkmal bei schwer zu erkennenden Zeichen darstellt, darf man aber auch keine „Wunder“ erwarten.



Spezielle, unbekannte Zeichen können trainiert werden um so die Erkennungsrate zu steigern.

Sprachdefinition

In einem weiteren Schritt werden die einzelnen ausgelesenen Zeichen zu ganzen Worten zusammengesetzt. Für diesen Prozess ist es wichtig, dass die OCR-Engine genau weiß, in welcher Sprache das Dokument verfasst wurde. Erst durch diese Voreinstellung lassen sich die Wortresultate auf ihre Sinnhaftigkeit hin überprüfen. Die Sprachdefinition liefert viele nützliche Zusatzinformationen für die Analyse der eingescannten Textvorlage. Wird z. B. ein englischsprachiges Dokument verarbeitet, lässt sich die Umlaut-Option, die für die englische Sprache nicht relevant ist, beim Analyseprozess von vornherein ausschließen. ABBYY OCR-Technologie verfügt darüber hinaus über die Fähigkeit, Dokumente zu erkennen, die mehrere Sprachen kombinieren.

Wörterbücher

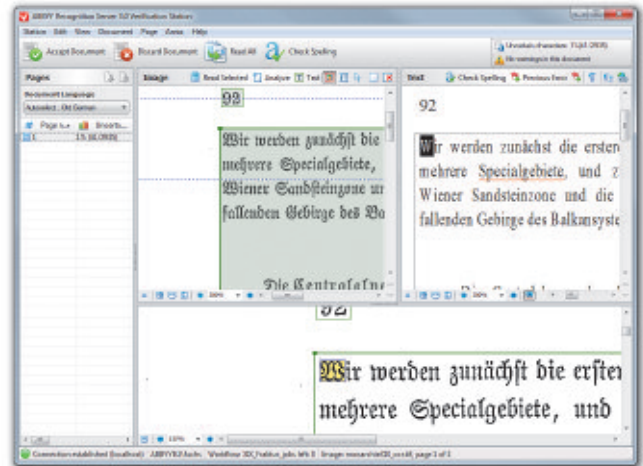
Um die Texterkennung auf Wortebene zu optimieren, machen sich ABBYY OCR-Produkte die Erkenntnisse neuester linguistische Technologien zunutze und verwenden im Hintergrund auch morphologische Wörterbücher als „Analysehelfer“. Dies ist wichtig, damit auch bei von der Software nur unsicher erkannten Zeichen, etwa bei einer sehr schlechten Dokumentenqualität der Textvorlage, eine möglichst richtige Textentscheidung getroffen werden kann. Ist beispielsweise ein „ü“ aufgrund schlechter Scan-Qualität nicht sicher zu erkennen, nutzt die Software weitere verfügbare Informationen, um den Text korrekt zu erfassen. Mit der Zusatzinformation „Es handelt sich um einen deutschen Text“ und einem Abgleich mit in der Software hinterlegten Wörterbüchern („München“ ist nicht im Wörterbuch zu finden, „München“ mit „ü“ hingegen doch), lässt sich der Text so trotz schlechter Vorlagequalität zuverlässig erkennen.

Was bei moderneren Texten hervorragend funktioniert, lässt sich allerdings nicht ohne weiteres auf die Erkennung historischer Texte übertragen. Denn hier fehlen sowohl genormte Standards für die Rechtschreibung, an denen sich eine OCR-Technologie orientieren könnte, als auch elektronisch verfügbare Wörterbücher aus älteren Zeiten. Um dennoch eine zuverlässige Texterkennung auch für schwierige historische Dokumente zu ermöglichen, ist es daher ratsam, gerade bei größeren Volumina und bestimmten Themenbereichen spezifische Wortlisten oder sogar ganze eigene Wörterbücher neu anzulegen. Natürlich lässt sich auch dieser zusätzliche Arbeitsprozess durch geeignete OCR-Produkte vereinfachen: Mit einem entsprechenden OCR-Entwicklungs-Toolkit können die Vorgänge für den Aufbau eines solchen Spezial-Wörterbuches über vorhandene Schnittstellen (APIs) entsprechend automatisiert werden.

Schritt 4 – Optionale Prüfung und Nachkorrektur durch den Anwender

Nach Durchführung der bisher beschriebenen Verarbeitungsschritte besteht zusätzlich die Möglichkeit, auch Anwender oder Operatoren in den Verarbeitungsprozess eingreifen zu lassen, um das Gesamtergebnis weiter zu verbessern. Gerade bei der Layoutanalyse (Definition der Text-, Bild- oder Tabellenblöcke)

kann eine manuelle Nachkorrektur sinnvoll sein. Eine manuelle Überprüfung kann auch zur Korrektur unsicher erkannter Zeichen und Wörter hinzugezogen werden.



Die Korrektur-Station von Recognition Server 3.0 erlaubt es die erkannten Texte vor dem Export zu überprüfen; Entwicklern und Nutzern der FineReader Engine (SDK) stehen hierzu sog. Visual Components zur Verfügung

Wichtiger Hinweis: In der Regel werden bei Massendigitalisierungsprojekten keine manuelle Korrektur angewendet, da dies bei mehreren hundert-tausend Seiten viel zu lange dauern würde. In der Praxis wird eher mit automatisierter, programmierbarer Nachkorrektur, basierend auf dem XML-Export gearbeitet. Dennoch kann in kritischen Fällen mit der Korrektur-Station direkt das Ergebnis „inspiziert“ und verifiziert werden.

Schritt 5 – Dokumentensynthese und Exportformate

In der Praxis erweist sich die Möglichkeit, unterschiedliche Ausgabeformate mit verschiedenen Optionen generieren zu können, als sehr wichtig, da sich das „eine“ universelle Format in der Regel nicht für alle nachgelagerten Arbeits- und Archivierungsschritte eignet. Gerade deshalb unterstützt ABBYY Technologie unterschiedlichste Exportformate: Ob als „Nur“-Text-, Office-, XML-Formate oder durchsuchbare PDF(/A)-Dokumente – Anwender können die eingescannten Textdokumente in den für sie benötigten digitalen Formaten exportieren¹.

Historische Dokumente können heute sowohl als Volltext, als durchsuchbare digitale Bücher oder als PDF-Dokumente exportiert werden. Ein XML-Export mit Zusatzinformationen (z. B. Zeichenkoordinaten) erlaubt es, die Ergebnisse in digitale Bibliotheken einzubinden und einem breiten Anwenderkreis zur Verfügung zu stellen. Die digitalisierten historischen Dokumente lassen sich darüber hinaus nicht mehr nur nach manuell erstellten Metadaten wie Autorennamen oder Buchtitel auffinden, sondern auch als Volltextdokumente nach bestimmten Suchbegriffen durchsuchen.

¹ Weitere Informationen zu verfügbaren Export-Formaten finden sich unter www.abbyy.de.

Entwicklung der ABBYY Fraktur-OCR

ABBYY arbeitet schon seit vielen Jahren eng mit Bibliotheken, Universitäten, Forschungsinstituten und anderen technologischen Partnern zusammen, um mit moderner OCR-Technologie auch Frakturschrift – eine Unterart der gotischen Schriften, wie sie typischerweise in Texten auftauchen, die zwischen 1800 und 1938 veröffentlicht wurden – automatisiert und zuverlässig zu erkennen. Als OCR-Spezialist stellte ABBYY seine jahrelange Erfahrung bei der Erkennung und digitalen Aufbereitung historischer Dokumente bereits in vielen nationalen und internationalen Projekten zur Digitalisierung von Bibliotheksbeständen zur Verfügung. Mit METAe und IMPACT war ABBYY auch an zwei wichtigen Forschungsprojekten der Europäischen Union zur Weiterentwicklung von Fraktur-OCR beteiligt. Um die technologischen Möglichkeiten zur automatisierten Erkennung von gebrochenen Schriften weiter zu verbessern, erstellten die Forschungsteams von ABBYY in dieser Zeit spezielle „Classifier“ oder Alphabete zur genauen Analyse von Frakturzeichen.

Dahinter verbirgt sich ein erheblicher Entwicklungsaufwand: Für jedes Frakturzeichen wurden durchschnittlich circa 2.500 Variationen hinterlegt, ein ganz neues Muster-Alphabet erstellt sowie 31.000 Seiten aus verschiedensten historischen Quellen gesammelt und detailliert getestet. Erst über die Verarbeitung einer Vielzahl von Beispieltextrn erlangte die ABBYY OCR-Engine die erforderliche Feinabstimmung, um auch Besonderheiten des Fraktur-Alphabets, wie z. B. Ligaturen oder zusammenhängende Buchstaben, automatisiert erkennen zu können.

METAe

Das EU-Forschungsprojekt METADATA ENGINE (METAe)² beschäftigte sich von 2000 bis 2003 mit der automatisierten Layout- und Strukturerkennung von Büchern und Zeitschriften sowie der Entwicklung einer OCR für Fraktur-Schriften. In einer Forschungsgemeinschaft von 14 Partnern aus Europa und den USA, darunter auch führende nationale Landesbibliotheken, erforschte das Projekt darüber hinaus mögliche Einsatzgebiete für die systematische Buch- und Zeitschriftendigitalisierung in Bibliotheken und Archiven.

Mit FineReader XIX entwickelte ABBYY für das METAe-Projekt eine spezielle Omnifont OCR-Lösung zur Erkennung von Frakturschrift. Diese Schriftart war in vielen europäischen Ländern weit verbreitet, im Fall der deutschen Sprache wurde sie sogar bis zum Jahr 1941 in etwa 80% aller gedruckten Dokumente verwendet. Die Omnifont-OCR-Lösung ABBYY FineReader XIX kann Fraktur ohne vorheriges individuelles „Training“ erkennen. Dafür erstellte ABBYY fünf historische Wörterbücher, die den historischen Stand der Rechtschreibung in den Sprachen Englisch, Französisch, Deutsch, Italienisch und Spanisch wiedergeben, zwischen 50.000 und 100.000 historische Wortstämme enthalten und damit über 90% der in historischen Texten vorkommenden Wörter abdecken³.

IMPACT

Auch im IMProving ACcess to Text (IMPACT)-Projekt⁴ nimmt ABBYY seit 2008 eine Schlüsselrolle wahr und unterstützt das Forschungsprojekt der EU-Kommission unter anderem durch die Bereitstellung modernster OCR-Technologien. Mit dem IMPACT-Projekt will die Europäische Union auch breiten Leserschichten außerhalb der Forschung den Zugang zu historischen Texten ermöglichen und die Digitalisierung des europäischen Kulturerbes durch die konsequente Weiterentwicklung und Verbesserung entsprechender Technologien vorantreiben. In enger Zusammenarbeit mit dem IMPACT-Projektteam verbesserten die Forschungs- und Entwicklungsteams von ABBYY in den letzten Jahren sowohl die Technologien für die Bildvorverarbeitung als auch diejenigen für die Analyse von Dokumentenlayouts. Unter Berücksichtigung unzähliger Beispieldokumente, die im Rahmen des Projektes von führenden europäischen Bibliotheken zur Verfügung gestellt werden, kümmert sich ABBYY um die Anpassung der OCR-Kernkomponenten für eine Vielfalt von historischen Druckdokumenten in verschiedenen europäischen Sprachen. Auf dieser Basis ließen sich bereits wichtige Technologiefortschritte für eine verbesserte Qualität der Frakturzeichenerkennung erzielen. Darüber hinaus entwickelte ABBYY im Rahmen des IMPACT-Projektes auch eine spezielle OCR-XML-Ausgabeoption zur Wiederherstellung logischer Dokumentenstrukturen.

² Weitere Informationen siehe <http://www.frakturschrift.com/de/projects/metae>

³ Ein Software Development Kit für FineReader XIX inklusive der Fraktur Classifier und historischen Wörterbücher ist über ABBYY erhältlich.

⁴ Weitere Informationen siehe <http://www.frakturschrift.com/de/projects/impact>

Warum sollte OCR auf alten Dokumenten angewendet werden?

OCR spielt eine wichtige Rolle bei der Digitalisierung und Entschlüsselung des in gedruckter Form vorliegenden Kulturerbes, und dies auf nationaler, europäischer Ebene als auch weltweit. Abseits überschaubarer Forscherzirkel besitzen die meisten Leser heutzutage heute nicht die Fähigkeit, Frakturschrift lesen zu können. Durch den Einsatz von Technologien wie Fraktur-OCR, die eine zuverlässige Texterkennung und umfassende Digitalisierung auch älterer Dokumente zunehmend möglich macht, lassen sich historische Quellen einer viel größeren Leserschicht als je zuvor zugänglich machen. Daher hat eine wie die von ABBYY entwickelte Omnifont-OCR, die auch ohne vorheriges Training Frakturschriften zuverlässig entziffert, eine große Bedeutung – nicht nur im Rahmen einzelner Digitalisierungsprojekte im Bibliotheksbereich, sondern für eine generelle Massendigitalisierung historischer Textdokumente.

Durch die heute immer einfacher werdende Anwendung von OCR können alte Texte nicht nur wiederentdeckt, sondern auch für aktuelle Nachdrucke verwendet werden. Besonders im wissenschaftlichen Bereich profitieren Anwender von der Digitalisierung. Durch die digitale Texterfassung können nun auch auf alten Dokumenten Suchfunktionen angewendet werden, was eine noch effizientere Erforschung der Unterlagen ermöglicht. Die mit OCR mögliche Umwandlung alter Texte in „moderne“ digitale Formate wie XML mit spezifischen Meta-Informationen (z. B. Informationen über das Original-Layout des Dokuments), durchsuchbare PDFs oder E-Books erweitert damit die Verbreitungsmöglichkeiten historischer Dokumente erheblich, ohne auf wertvolle oder möglicherweise bereits gefährdete Papieroriginale zurückgreifen zu müssen.

Qualität bei „historischer“ OCR-basierter Texterkennung

Obwohl sich viele Projekte der kontinuierlichen Weiterentwicklung der Technologien zur Frakturerkennung widmen und, wie oben geschildert, auch bereits erhebliche Fortschritte in den letzten Jahren erzielt wurden, stellt sich trotzdem immer wieder die Frage nach der Erkennungsqualität, denn: Wann ist „gut“ wirklich „gut genug“?

Auch wenn sich das Thema hier fast schon im Bereich einer philosophischen Fragestellung bewegt, sollen dennoch einige qualitätsrelevante Aspekte kurz aufgezeigt werden:

Die erzielbare Gesamtgenauigkeit hängt generell von vielen verschiedenen Parametern ab, wie in der folgenden Übersicht kurz skizziert:

- Papierqualität des Originals
- Qualität des Scans
- Richtige Scan-Parameter
- Qualität der Bildvorverarbeitung
- Qualität der Dokumentenanalyse zur genauen Identifikation aller im Dokument vorhandenen Text- und Bildbereiche
- Genaue Rekonstruktion des Layouts
- Einhaltung der Lese-Reihenfolge
- Optimierte Zeichenerkennung für Antiqua Schriften
- Einsatz einer OCR-Lösung mit Spezialisierung auf Frakturschrift
- Verfügbarkeit geeigneter Wörterbücher
- Möglichkeiten zur manuellen und/oder automatisierten (Nach-) Korrektur

Natürlich besteht in der Praxis immer der Wunsch nach einer fast 100-prozentig genauen Zeichenerkennung. Wie oben geschildert lässt sich das aufgrund der spezifischen Besonderheiten historischer Textdokumente nicht immer automatisch erreichen. Die sehr hohen Qualitätsstandards, die sich bei der OCR-Anwendung auf modernen Dokumenten heute sehr einfach erreichen lassen, kosten beim Umgang mit historischem Material wesent-

lich mehr Zeit und sind daher auch sehr viel kostenaufwändiger – sowohl mit Blick auf die Vorbereitungsprozesse, die Projektdurchführung selbst sowie die manuelle oder automatische Nachbearbeitung der Texterkennungsresultate.

Bei der Wahl zwischen „nur“ zehn in sehr hoher Qualität erschlossenen Büchern oder 1000 digital erschlossenen historischen Dokumenten würden viele Anwender allerdings wohl eher Letzteres bevorzugen. In der Praxis muss daher immer sehr genau abgewogen werden, welcher Aufwand für welchen Anwendungszweck betrieben werden kann oder sollte.

Da historische Dokumente naturgemäß auch in ihrem historischen Kontext verhaftet sind, das heißt sowohl in einer unmodernen Sprache als auch in einer historischen Rechtschreibung verfasst wurden, die sich teilweise drastisch von aktuellen Schreibweisen unterscheidet, ist der Einsatz intelligenter Suchtechnologien bei der Erschließung dieser Dokumente in jedem Fall notwendig, z.B.

- alte Schreibweisen: Theile, Mittheilung, reduziert;
- moderne Schreibweisen: Teile, Mitteilung, reduziert.

Die dabei verwendeten Suchtechnologien sind in der Regel auch in der Lage, Wörter mit OCR-Fehlern dennoch zu finden. Der weitere Vorteil eines in moderner Schrift vorliegenden historischen Dokumentes liegt auf der Hand: Auch der Frakturschrift unkundige Leser können damit historische Quellen einfacher lesen und nutzen.

Bei Lesern, die heute digitale Bibliotheken nutzen und sich für historische Dokumente interessieren, handelt es sich in der Regel um interessierte und mündige Leser, die die Inhalte trotz eventueller Erkennungsungenauigkeiten richtig verstehen und interpretieren können. Die Zusatzinformation, dass es sich bei einem Text um ein automatisch OCR-erfasstes Dokument ohne manuelle Nachbearbeitung handelt, wird jeden Leser in die Lage versetzen, diesen Umstand richtig zu deuten.

Fazit

Bei Fraktur-OCR handelt es sich immer um relativ komplexe Projekte und gerade bei der automatisierten Erkennung historischer Dokumente wird sicherlich kein Projekt dem anderen gleichen. Unterschiede bei der Qualität der Vorlagen, den Layouts, den verwendeten Schriften oder auch fehlende Lexika legen die Hürden sehr hoch.

Trotzdem erlauben die aktuell erreichten Optimierungen bei der Verarbeitung historischer Dokumente schon heute, OCR auch auf bereits gescannte Bildbestände historischer Dokumente anzuwenden. Sowohl die benötigte Rechenleistung als auch eventuell entstehende Lizenzkosten bei der Anwendung von OCR-Lösungen erweisen sich längst nicht mehr als „K.O.-Kriterium“, um nicht auch alten, bisher nur in gedruckter Form vorliegenden Bibliotheksbeständen die Chance auf ein „zweites“, digitales Leben zu geben⁵.

Bereits heute deuten sich auch ganz neue Möglichkeiten zur weiteren Erschließung und Korrektur digitaler Dokumente an. So bietet sich etwa die zunehmende Verbreitung offener Systeme im Umfeld des sogenannten „Crowd Sourcing“ – der Korrektur von Texten durch Freiwillige, die digitalisierte Bücher querlesen und verbessern – als vielversprechende Option an, damit alte Texte und Quellen auch in Zukunft ihre wichtige Rolle für das Verstehen von Gegenwart und Vergangenheit nicht verlieren.

ABBYY®

ABBYY Europe GmbH

Eisenheimerstrasse 49
80687 Munich, Germany
Tel: +49 89 511 159 0
Fax: +49 89 511 159 59
sales_eu@abbyy.com
www.ABBYY.com
www.ABBYY.de

Bureau France
4, rue Leroux
94100 Saint-Maur des Fossés
France
sales_france@abbyy.com
www.france.ABBYY.com

ABBYY UK Ltd.

Abbey House, Grenville Place
Bracknell RG12 1BP, United Kingdom
Tel: +44 1344 392 610
Fax: +44 1344 392 611
sales_UK@abbyy.com
www.ABBYY.com

⁵ Weitere Informationen und Links rund um Fraktur OCR unter www.frakturschrift.de.