ABBYY®

FALLSTUDIE

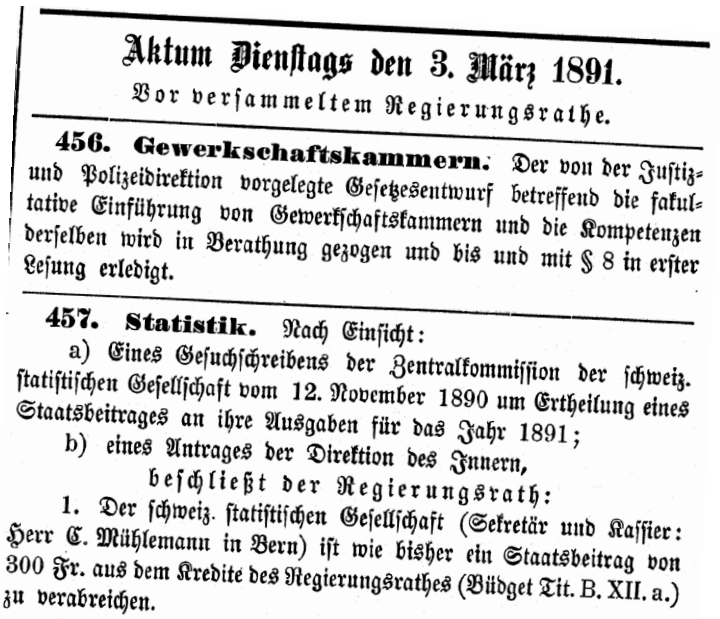
ABBYY® Historic OCR

Digitaler Zugang zur Vergangenheit

Das Staatsarchiv Zürich und das Institut für Computerlinguistik der Universität Zürich konvertieren mit ABBYY Fraktur OCR 11.000 Seiten an Regierungsratsbeschlüssen, um diese öffentlich online zugänglich zu machen.

Die Erschließung von Kulturgut durch die Digitalisierung schriftlicher Quellen ist ein wichtiger Schritt zu einer besseren informellen Infrastruktur, die sowohl der Forschung, als auch dem Wissenstand der breiten Masse zu Gute kommt. Durch ein starkes Bewusstsein für den gesellschaftlichen Wert kulturellen Gutes ist ABBYY seit dem Jahr 2000 in verschiede Projekte involviert, die sich sowohl mit der Relevanz des Erhalts und der Zugänglichkeit kulturellen Erbes sowie mit der Weiterentwicklung der benötigten Technologien befassen. Kulturwissenschaftliche Forschung ist seit jeher bedingt durch den Stand der Technologien, die ihr zur Verfügung stehen. Gleichzeig treibt der technische Bedarf kulturwissenschaftlicher Projekte die technologische Weiterentwicklung voran. Als wissenschaftliche Kooperation zwischen dem Staatsarchiv Zürich und dem Institut für Computerlinguistik der Universität Zürich ist das Projekt „RRB-Fraktur“ ein beispielhafter Nachweis für das enge Zusammenspiel kulturwissenschaftlicher und technologischer Interessen. Basierend auf ABBYYs Fraktur OCR-Technologien ist es Ziel des Projekts sowohl Kulturgut zur Verfügung zu stellen, als auch den Stand der OCR-Wissenschaft voranzutreiben.

Um mehr staatliche Dokumente für den öffentlichen Gebrauch online bereit zu stellen, und somit sowohl seinen Bürgern als auch der Forschung einen Zugang zu informativem Material zu bieten, entschied sich die Regierung des Kantons Zürich Regierungsratsbeschlüsse („RRB“) aus den Jahren 1803 bis 1995 zu digitalisieren. Darunter sind 11.000 Seiten aus den 16 Jahrgängen von 1887 bis 1902, die großteils in Fraktur gedruckt sind. Im wissenschaftlichen Interesse steht dabei nicht nur die Bereitstellung noch unerforschter geschichtlicher Quellen, sondern vor allem auch die technologische Herausforderung des Projekts: Da die Dokumente nicht nur in Frakturschrift, sondern teilweise in einer Mischung aus Fraktur und „normalen“ Antiqua-Anteilen vorliegen, wird die optische Zeichenerkennung (OCR) vor eine schwierige Aufgabe gestellt. Im Vordergrund stehen dabei nicht nur die Erkennung des Textes, sondern vor allem auch die Optimierung der OCR-Nachbearbeitung und das Ausschöpfen der Möglichkeiten von OCR.



Aktum Dienstags den 3. März 1891.

Vor versammeltem Regierungsrathe.

456. Gewerkschaftskammern. Der von der Justiz- und Polizeidirektion vorgelegte Gesetzesentwurf betreffend die fakultative Einführung von Gewerkschaftskammern und die Kompetenzen derselben wird in Berathung gezogen und bis und mit § 8 in erster Lesung erledigt.

457. Statistik. Nach Einsicht:

a) Eines Gesuchschreibens der Zentralkommission der schweiz. statistischen Gesellschaft vom 12. November 1890 um Ertheilung eines Staatsbeitrages an ihre Ausgaben für das Jahr 1891;

b) eines Antrages der Direktion des Innern,

beschließt der Regierungsrath:

1. Der schweiz. statistischen Gesellschaft (Sekretär und Kassier: Herr C. Mühlemann in Bern) ist wie bisher ein Staatsbeitrag von 300 Fr. aus dem Kredite des Regierungsrathes (Budget Tit. B. XII. a.) zu verabreichen.

Über das Staatsarchiv Zürich

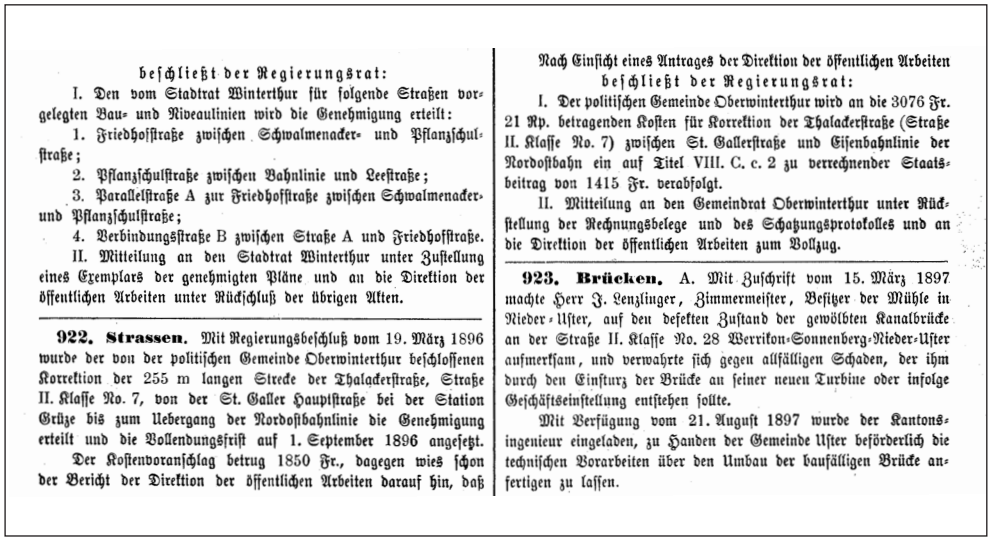
Das Staatsarchiv ist das Archiv der öffentlichen Organe des Kantons Zürich, welche den Kantonsrat, die Regierung, die kantonalen Zentral- und Bezirksverwaltung sowie die Gerichte und Anstalten umfassen. Es übernimmt, erschließt und konserviert deren überlieferungswürdige Unterlagen. Als historisches Archiv verwahrt das Staatsarchiv zudem das Verwaltungsschriftgut des alten Stadtstaates Zürich seit der Zeit des Mittelalters. Ergänzt werden diese Bestände durch Dokumente privater Herkunft (z. B. von Firmen, Vereinen, Zünften, Familien und Einzelpersonen). Die Aufbewahrung dieser Unterlagen soll staatliches Handeln nachvollziehbar machen, historische Forschungen ermöglichen und kulturelle Interessen im weitesten Sinn bedienen.

Kontakt

Staatsarchiv des Kantons Zürich
Winterthurerstrasse 170
8057 Zürich
Telefon +41 44 635 69 11
Fax +41 44 635 69 05

ABBYY Europe GmbH
Elsenheimerstr. 49, D-80687 München
Tel: +49 89 51 11 59 - 0, Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com

OCR an sich ist bereits ein komplexer informatischer Vorgang. Die Erkennung Frakturschriften stellt eine besondere Herausforderung dar, an der herkömmliche OCR-Software scheitert. Neben der häufig schlechten Papierqualität der historischen Vorlagen existierten beim Druck älterer Druckdokumente oft keine standardisierten Schriften. Daher kann das Aussehen gleicher Buchstaben von Dokument zu Dokument stark voneinander abweichen. Zudem haben sich Sprache und Rechtschreibung in den letzen Jahrhunderten stark verändert, sodass aktuelle Wörterbücher in der Regel nicht zu einer automatischen Korrektur herangezogen werden können.



beschließt der Regierungsrat:

I. Den vom Stadtrat Winterthur für folgende Straßen vorgelegten Bau- und Niveaulinien wird die Genehmigung erteilt:

1. Friedhofstraße zwischen Schwalmenacker- und Pflanzschulstraße;
2. Pflanzschulstraße zwischen Bahnlinie und Leestraße;
3. Parallelstraße A zur Friedhofstraße zwischen Schwalmenacker- und Pflanzschulstraße;
4. Verbindungsstraße B zwischen Straße A und Friedhofstraße.

II. Mitteilung an den Stadtrat Winterthur unter Zustellung eines Exemplars der genehmigten Pläne und an die Direktion der öffentlichen Arbeiten unter Rückschluß der übrigen Akten.

922. Strassen. Mit Regierungsbeschluß vom 19. März 1896 wurde der von der politischen Gemeinde Oberwinterthur beschlossenen Korrektion der 255 m langen Strecke der Thalackerstraße, Straße II. Klasse No. 7, von der St. Galler Hauptstraße bei der Station Grüze bis zum Uebergang der Nordostbahnlinie die Genehmigung erteilt und die Vollendungsfrist auf 1. September 1896 angesetzt.

Der Kostenvoranschlag betrug 1850 Fr., dagegen wies schon der Bericht der Direktion der öffentlichen Arbeiten darauf hin, daß

Nach Einsicht eines Antrages der Direktion der öffentlichen Arbeiten beschließt der Regierungsrat:

I. Der politischen Gemeinde Oberwinterthur wird an die 3076 Fr. 21 Rp. betragenden Kosten für Korrektion der Thalackerstraße (Straße II. Klasse No. 7) zwischen St. Gallerstraße und Eisenbahnlinie der Nordostbahn ein auf Titel VIII. C. c. 2 zu verrechnender Staatsbeitrag von 1415 Fr. verabfolgt.

II. Mitteilung an den Gemeindrat Oberwinterthur unter Rückstellung der Rechnungsbelege und des Schatzungsprotokolles und an die Direktion der öffentlichen Arbeiten zum Vollzug.

923. Brücken. A. Mit Zuschrift vom 15. März 1897 machte Herr J. Lenzlinger, Zimmermeister, Besitzer der Mühle in Nieder-Uster, auf den defekten Zustand der gewölbten Kanalbrücke an der Straße II. Klasse No. 28 Werrikon-Sonnenberg-Nieder-Uster aufmerksam, und verwahrte sich gegen allfälligen Schaden, der ihm durch den Einsturz der Brücke an seiner neuen Turbine oder infolge Geschäftseinstellung entstehen sollte.

Mit Verfügung vom 21. August 1897 wurde der Kantonsingenieur eingeladen, zu Handen der Gemeinde Uster beförderlich die technischen Vorarbeiten über den Umbau der baufälligen Brücke anfertigen zu lassen.

Als führender Hersteller von OCR-Technologien ist ABBYY auch an der Entwicklung von Fraktur OCR beteiligt. Mit dem Ziel bestehende Technologien und Produkte kontinuierlich weiterzuentwickeln, beteiligte sich ABBYY an dem Forschungsprojekt (2008 bis 2011) der EU-Kommission IMProving ACcess to Text (IMPACT). ABBYY lieferte nicht nur die OCR-Basistechnologien, sondern wirkte aktiv in enger Zusammenarbeit mit den anderen Projektteams an der Verbesserung der Erkennungstechnologien mit.

Die Verarbeitung der Regierungsratsbeschlüsse wurde ebenfalls mit ABBYY-Technologie umgesetzt: Erste Testläufe wurden mit ABBYY FineReader XIX durchgeführt. Diese Fraktur OCR ist seit 2004 auf dem Markt und lieferte zufriedenstellende Ergebnisse. Jedoch ergaben sich bei dieser alten Version Probleme, wenn „normale“ Antiqua- und Frakturschrift abwechselnd im Text auftraten. Daher entschied sich das Forschungsteam ABBYY Recognition Server 3.0 zu verwenden, da die Software Dokumente mit gemischten Schriftarten erkennen kann und auch weitere Vorteile und Verbesserungen aus dem IMPACT-Projekt bietet. Zudem ist Recognition Server in der Lage alle Text- und Layout-Informationen in einer XML-Datei auszugeben, wodurch sich die Lösung besonders für Forschung und automatisierte Nachbearbeitung (Post-Correction) eignet. Da XML detaillierte Informationen über die Koordinaten eines Schriftzeichens im Dokument, die Korrektheit des Wortes, die Erkennungs- und Fehlerwahrscheinlichkeit enthält, ist das Ausgabeformat die optimale Basis für die vom Institut für Computerlinguistik der Universität Zürich angestrebte Nachkorrektur des OCR-Ergebnisses.

Die durchschnittliche Erkennungsgenauigkeit der Fraktur-Dokumente lag bei 97,2% auf Wortebene. Diese Genauigkeit ist für viele Zwecke ausreichend, z. B. für die Textsuche, Lesbarkeit oder Recherche. Das Institut für Computerlinguistik strebte jedoch eine Verbesserung der Erkennungsqualität durch eine intelligente Nachkorrektur an. Ansatzpunkte hierzu waren die Nutzung externer, spezialisierter Wörterbücher und Informationen, die aus dem Text selbst stammen.

Das Team um Prof. Dr. Martin Volk konzentrierte sich auf die Optimierung der systematischen, wiederkehrenden OCR-Fehler, d.h. eine möglichst genaue Einstufung des erkannten Wortes als „falsch“ oder „richtig“. Obwohl Recognition Server bereits eine sehr gute Erkennungsquote aufwies, waren die Informationen bezüglich der Erkennungswahrscheinlichkeit nicht immer zuverlässig. Das im OCR-Prozess genutzte Wörterbuch stufte teilweise auch fehlerhaft erkannte Wörter als „richtig“ ein und korrekt erkannte Wörter wurden als „unsicher“ markiert, da sie nicht im Wörterbuch enthalten waren.

Über die Universität Zürich

Die Universität Zürich nimmt als größte Universität der Schweiz eine herausragende Stellung in der Forschungs- und Bildungslandschaft des Landes ein. Sie ist höchsten internationalen Wissenschaftsstandards und verantwortungsvoller Reflexion verpflichtet. Als Mitglied der «League of European Research Universities» (LERU) gehört sie zum Kreis der besten Europäischen Forschungsuniversitäten und ist höchsten internationalen Wissenschaftsstandards und verantwortungsvoller Reflexion verpflichtet. Mit rund 100 Fächern verfügt die Universität Zürich über die größte Vielfalt im Studienangebot in der Schweiz. Die Universität Zürich erbringt wissenschaftliche Dienstleistungen für die Wirtschaft und Gesellschaft indem sie ihr Wissen zielgruppengerecht und abwechslungsreich an die interessierte Öffentlichkeit weitergibt. Die Universität Zürich fördert außerdem den Wissenstransfer in die Wirtschaft und schafft durch Kollaborationen attraktive Arbeitsplätze in zukunftsgerichteten Wirtschaftsbereichen.

Kontakt

Universität Zürich
Rämistrasse 71
CH-8006 Zürich
Tel. +41 44 634 11 11
Fax +41 44 634 49 01

ABBYY Europe GmbH
Elsenheimerstr. 49, D-80687 München
Tel: +49 89 51 11 59 - 0, Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com

Da bei einer bestimmten Gruppe von Texten häufig ähnliche OCR-Fehler bei verschiedenen Instanzen des Wortes auftreten, konnte eine Verbesserung der Erkennungsquote durch eine Spezifikation des verwendeten Wörterbuches erreicht werden. Dies traf auch auf das „RBB"-Projekt zu, da in staatlichen Dokumenten aus einer bestimmten Region die Sprachwahl oft sehr einheitlich ist. Durch eine entsprechende Anpassung des Wörterbuches mit ausführlicheren und mehr Einträgen, wie z. B. Ortsnamen, regionsspezifischen Schreibweisen und Fachjargon, konnte die Erkennungsrate gesteigert werden. Das daraus resultierende Korpus-Lexikon, das morphologische Variationen und Zusammensetzungen sowie lokale Toponyme (Ortsnamen) enthielt und historische sowie regionale Abweichungen der Orthographie berücksichtigt, gleicht nach den Prinzipien „Häufigkeit" und „Morphologie" die von Recognition Server eingestufte Fehleranfälligkeitswahrscheinlichkeit ab.

Obwohl die Integration eines Korpus-Lexikons die Erkennungsquote verbesserte, blieben dennoch seltene Wörter nach wie vor unerkannt. 50% der OCR-Fehler haben ihre Ursache darin, dass einzelne Buchstaben nicht richtig erkannt werden. Durch das systematische Austauschen und Abgleichen ähnlicher Buchstaben aus Listen, in denen Buchstaben und häufige Buchstabenverwechslungen manuell in einer Ersetzungstabelle zusammengestellt werden, kann das OCR-Ergebnis weiter verbessert werden. Wird ein Wort bis dahin immer noch nicht erkannt, wird die so genannte N-Gramm-Analyse (n-gram hashing) angewandt, um nach ähnlichen Wörtern zu suchen. Die Wörter werden in sog. Trigramme eingeteilt und anhand Levenshtein'schen Editierdistanz in eine bestimmte Erkennungswahrscheinlichkeitskategorie eingeteilt.

Um letztendlich die fast 100%ige Richtigkeit des erkannten Textes garantieren zu können, bleibt jedoch nur die manuelle Überprüfung und Korrektur des Textes durch einen Lektor. Durch die Kosten für die Arbeitszeit ist dies bei umfangreichen Digitalisierungsprojekten in der Regel keine Option. Um deshalb auch diesen Aufwand dennoch so zeit- und kosteneffizient wie möglich zu halten, greifen schon einige Projekte auf „Crowd Correction" zurück. Dieses „Wikipedia-Prinzip" bietet einen einfachen öffentlichen Zugang, um so die Korrektur eines Textes zu ermöglichen. Die Ergebnisse werden kostenlos online zur Verfügung gestellt und dann von Nutzern des Archivs überprüft und korrigiert.

Das Projekt RRB-Fraktur wird mittlerweile erfolgreich umgesetzt. Zwar ist Fraktur OCR nicht für jedermann notwendig, doch wäre dieses Projekt ohne ABBYYs Technologie nicht möglich gewesen, da gängige OCR-Software an den Herausforderungen von Fraktur scheitert. ABBYY Recognition Server 3.0 lieferte die optimale Basis für die Entwickler des Projekts, da die Lösung bereits von Anfang an sehr gute Erkennungsergebnisse darbot. Zudem erlaubt der XML-Export der Universität wissenschaftlich weitere Optimierungsansätze zu testen und auch praktisch im Projekt umzusetzen.

„Wir sind sehr zufrieden mit den Fortschritten, die wir mit Hilfe von ABBYY erreichen konnten", bestätigt Lenz Furrer, Mitarbeiter des Instituts für Computerlinguistik der Universität Zürich. „Wir werden auch in zukünftigen Projekten von den Erkenntnissen und Ergebnissen, die wir durch das Projekt erlangt haben, profitieren."

Wer mehr über die technischen Details erfahren möchte, findet unter www.frakturschrift.com unter anderem ein PDF der wissenschaftlichen Veröffentlichung des Forschungsteams von Prof. Dr. Martin Volk.

Über ABBYY

ABBYY ist ein führendes Unternehmen in der Entwicklung von Technologien für Dokumenterkennung, Dokumentenumwandlung, Data Capture und Linguistik.

Die Abteilung Forschung & Entwicklung widmet sich der kontinuierlichen Weiterentwicklung und Verbesserung der OCR-Technologien, um eine noch genauere Erkennung einer immer breiteren Masse von Dokumenten zu erreichen. Seit dem Jahr 2003 ist ABBYY in die Entwicklung von Fraktur OCR involviert, einer speziellen Technologie für die Erkennung von gebrochenen Schriften. Kunden und Partner profitieren von ABBYYs Forschungsarbeit durch die Verfügbarkeit der Technologieentwicklungen in den neuesten Produktversionen.

Zum Produktportfolio von ABBYY gehören: ***FineReader OCR*** *und* ***PDF Transformer*** *– Endanwenderprogramme zur Umwandlung von Dokumenten;* ***Recognition Server*** *– eine serverbasierte Lösung für OCR und PDF-Umwandlung;* ***FlexiCapture*** *– Data Capture Lösung zur Verarbeitung von Formularen, semi- und unstrukturierten Dokumenten;* ***FineReader Engine SDKs*** *mit dem gesamten Leistungsumfang der ABBYY OCR-Technologien;* ***Lingvo*** *– eine Serie von elektronischen Wörterbüchern.*

Mehr Informationen über ABBYY unter www.ABBYY.com

ABBYY®

ABBYY Europe GmbH
Elsenheimerstr. 49, D-80687 München
Tel: +49 89 51 11 59 - 0, Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com