Vom Papier zur Datenanalyse. „Neue“ historische Forschungsdaten für die Wirtschaftswissenschaften

Sabine Gehrlein, Jan Kamlah, Matthias Pintsch, Irene Schumm und Stefan Weil
Universitätsbibliothek Mannheim, Deutschland

1. Ausgangslage

Die empirische Wirtschaftsforschung hat eine lange Tradition und damit auch die Erhebung und Bereitstellung der benötigten Forschungsdaten. Neben den Akteuren der amtlichen Statistik hat sich ein differenzierter und internationaler Kreis von kommerziellen Datenanbietern etabliert, die ihre Produkte hochpreisig anbieten. Zu den Schwergewichten in diesem Markt gehören etwa Bloomberg, Bureau van Dijk, Thomson Reuter's, Standard & Poor's oder Bisnode. Dabei ist die Wissenschaft nicht die Hauptzielgruppe dieser Unternehmen; im Fokus stehen eher große Finanz- und Versicherungsunternehmen und deren Marktforschungsabteilungen. Ein Blick in das Datenbank-Infosystem (DBIS) zeigt, dass aktuell nur eine kleine Zahl deutscher Hochschulen die notwendigen Mittel aufbringen kann, um ihren Forscherinnen und Forschern Zugang zu den Produkten der kommerziellen Datenanbieter zu verschaffen.

Selbst wenn dies der Fall ist, ist damit noch keine umfassende Versorgung der empirischen wirtschaftswissenschaftlichen Forschung gewährleistet. Die verfügbaren kommerziellen Forschungsdaten sind geographisch recht einseitig ausgerichtet: Insbesondere für die USA existiert ein großes und differenziertes Angebot, für Deutschland hingegen sind deutlich weniger Ressourcen verfügbar.

Zudem lassen sich viele Fragestellungen nur dann methodisch sauber bearbeiten, wenn die Forschungsdaten über einen ausreichend langen Zeitraum konsistent verfügbar sind. Diese langen Zeitreihen können die kommerziellen Datenanbieter nur selten liefern. Da die primären Kunden in der Finanz- und Versicherungsbranche vor allem an aktuellen Daten und Analysen interessiert sind, wird in die Pflege und Aufbereitung historischer Daten kaum investiert. Abbildung 1 zeigt die zeitliche und inhaltliche Abdeckung von drei zentralen kommerziellen Unternehmensdatenbanken, aus der anschaulich hervorgeht, dass historische Phänomene wie z. B. die Ölkrise der 70er Jahre auf dieser Datengrundlage nicht analysiert werden können.

In Zukunft könnte sich diese Situation sogar noch verschärfen, da viele relevante Wirtschaftsdaten, die online über Webdienste angeboten werden, verschwinden, sobald sie mit der aktuellen Version überschrieben werden. Diese Daten sind dann für die Forschung verloren.

Hier können nicht-kommerzielle Anbieter, darunter auch Bibliotheken, eine wichtige Rolle spielen (vgl. Zumstein 2016). Mit der Expertise, die sie in Fragen des Datenmana-

Sabine Gehrlein et al.: Vom Papier zur Datenanalyse. „Neue“ historische Forschungsdaten für die Wirtschaftswissenschaften.
In: Heuveline, V., Gebhart, F. und Mohammadianbisheh, N. (Hrsg.): E-Science-Tage 2019 – Data to Knowledge,
Heidelberg: heiBOOKS, 2020, 140–152. DOI: https://doi.org/10.11588/heibooks.598.c8423

	COMPUSTAT GLOBAL	EIKON /DATASTREAM	ORBIS / AMADEUS
Anbieter	Standard & Poor's	Thomson Reuters	Bureau van Dijk
Zeitraum	1988-heute	Ø Letzte 5 Jahre	Letzte 10 Jahre
Anzahl dt. Firmen	942	ca. 1.000	3,85 Mio.
Bilanzdaten	Ja	Ja	Ja
Governancedaten	Nein	Ja	Ja

Abbildung 1.: Übersicht über die Inhalte gängiger kommerzieller Forschungsdatenangebote (eigene Darstellung)

gements, der Erschließung und der Archivierung elektronischer Informationen aufgebaut haben, sind sie geeignete Akteure, um alte und neue Forschungsdaten zu sammeln, aufzubereiten und dauerhaft für die Wissenschaft verfügbar zu machen.

Das Angebot muss dabei an den aktuellen Bedürfnissen der wissenschaftlichen Nutzerinnen und Nutzer ausgerichtet werden. So ist es in der quantitativen Wirtschaftsforschung heute Standard, mit Auswertungs- und Analyseprogrammen wie Stata, R oder SPSS in kurzer Zeit sehr große Datenmengen zu verarbeiten und so Theorien und Hypothesen zu testen. Fragen, die früher langwierige Datensammlungen und Vorbereitungen voraussetzten, lassen sich so innerhalb weniger Stunden bearbeiten.

Voraussetzung dafür ist, dass Rohdaten in einem maschinell zu verarbeitenden Format und in hoher inhaltlicher Qualität vorliegen. Es reicht daher nicht aus, eine Datenquelle nur zu digitalisieren. Zusätzlich ist mindestens eine nutzbare Volltexterfassung über OCR nötig, besser noch die Überführung der Daten in eine strukturierte Datenbank mit vielfältigen Such- und Downloadmöglichkeiten. Zudem sollten die angebotenen Daten inhaltlich umfassend beschrieben und dokumentiert sowie einfach und kostengünstig zugänglich sein.

2. Das Projekt Aktienführer an der UB Mannheim

Die Universität Mannheim hat ein ausgeprägtes Profil in den empirisch orientierten Wirtschafts- und Sozialwissenschaften, was sich auch in national und international renommierten Fachbereichen für Betriebswirtschaftslehre und Volkswirtschaftslehre widerspiegelt. Voraussetzungen dafür sind insbesondere erfolgreiche Forschungsprojekte und daraus resultierende Publikationen, die auch auf einem erstklassigen Zugang zu Publikationen und Datenbanken beruhen. Dienstleister und Ansprechpartner in diesem Feld ist die Universitätsbibliothek (UB).

Im alltäglichen Austausch mit Forscherinnen und Forschern der Betriebswirtschaftslehre erhielt die UB wiederholt Hinweise darauf, dass es der Forschung am Zugang zu Forschungsdaten im Bereich der deutschen Unternehmens- und Kapitalmarktdaten mangelt. Konkret handelte es sich um Daten aus dem Hoppenstedt-Aktienführer, einer Publikation, die mit ihrem Vorgänger Saling's Börsenpapiere seit dem Jahr 1870 jährlich zu den deut-

schen börsennotierten Unternehmen berichtete. In jährlichen Print-Bänden und ab 2000 auf CD-ROM enthält der Hoppenstedt-Aktienführer detaillierte Daten zur Entwicklung wichtiger Unternehmen und der Volkswirtschaft Deutschlands.

Vor dem Hintergrund eines deutlich kommunizierten Bedarfs von Seiten der Forschung setzte sich die UB Mannheim das Ziel, die wertvollen Daten des Hoppenstedt-Aktienführers in ein zeitgemäßes, elektronisches Format zu überführen und der Forschung zur Verfügung zu stellen. Dieses Projekt erschien naheliegend, da die UB Mannheim aufgrund des Universitätsprofils bestens mit wirtschaftswissenschaftlichen Forschungsdaten vertraut ist und da vorhandene Kompetenzen im Feld der automatisierten Texterkennung (OCR – Optical Character Recognition) eingebracht und weiterentwickelt werden konnten. Diese Kompetenzen spielen auch im Rahmen weiterer Projekte eine große Rolle.[1]

Durch Förderung der DFG war es der UB Mannheim möglich, den Hoppenstedt-Aktienführer im Rahmen zweier Projekte von 2013 bis 2015 sowie von 2017 bis 2019 zu digitalisieren, in eine komfortable Datenbank zu überführen und der Forschung kostenfrei zur Verfügung zu stellen. Durchführung und Ergebnis, insbesondere der zweiten Projektphase, sollen weiter unten genauer diskutiert werden.

In der deutschen Unternehmens- und Kapitalmarktforschung ist der Hoppenstedt-Aktienführer (kurz: Aktienführer) eine bestens etablierte Datenquelle. Bevor der Hoppenstedt-Verlag die Publikation übernahm, erschien sie im Verlag Saling – weswegen manche den Aktienführer auch noch als „Salinger“ kennen. Derzeit liegen die Nutzungsrechte beim Unternehmen Bisnode, von dem die UB Mannheim die Rechte zur Digitalisierung und elektronischen Präsentation erworben hat.

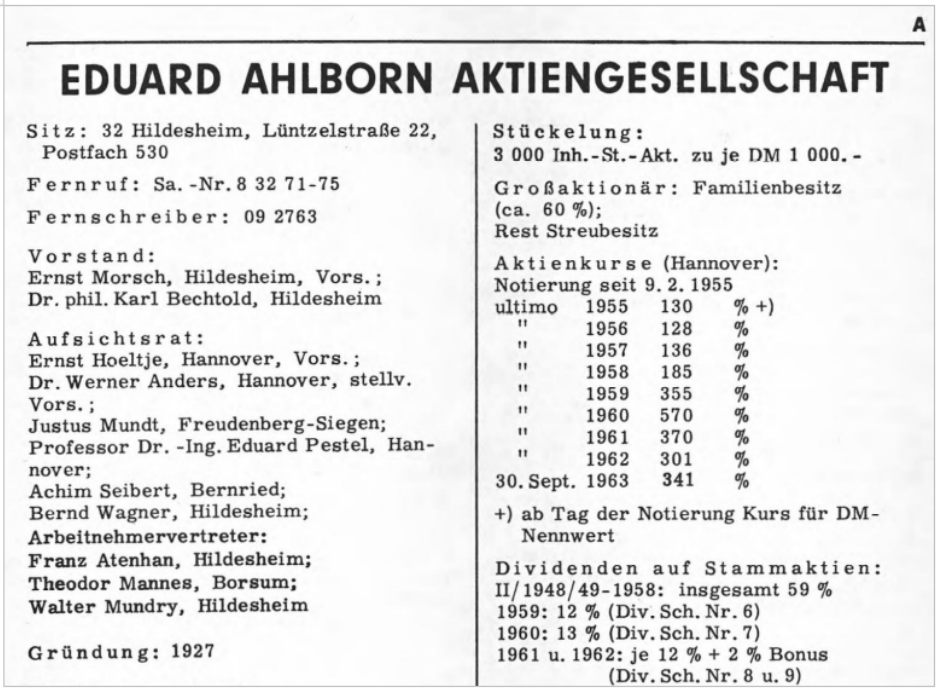
Nach dem Zweiten Weltkrieg erschien der Aktienführer erstmals wieder 1953 und dann ab 1956 in jährlich aktualisierten Print-Ausgaben, die 1999 eingestellt wurden. Ab dem Jahr 1998 und bis 2018 gab es dann eine halbjährlich aktualisierte CD-ROM, und seit 2019 bietet Bisnode die Aktienführer-Daten nur noch elektronisch als Web-Dienst an. Es ist noch zu klären, wie die Daten des Web-Dienstes für die Forschung gesichert und dauerhaft verfügbar gemacht werden können.

Die seit 1956 erschienenen Aktienführer-Bände und CD-ROMs enthalten in standardisierter Form Berichte über deutsche und ausländische Aktiengesellschaften, die an einer deutschen Börse gehandelt werden. In den Berichten finden sich unter anderem Daten zu Firmensitz und Vorstand, zu den Tätigkeitsbereichen und Beteiligungen, zur Aktionärs- und Kapitalstruktur und zu wesentlichen Positionen der Bilanz sowie Gewinn- und Verlustrechnung. Die konstant hohe Datenqualität sowie die lange Publikationshistorie des Aktienführers machen die Daten sehr wertvoll für die auf Deutschland bezogene empirische Wirtschaftsforschung. Abbildung 2 zeigt beispielhaft einen Ausschnitt aus einem Unternehmensprofil.

Durch die stringente Struktur der Unternehmensprofile in diesen Aktienführer-Bänden war es möglich, die OCR-Erkennung und Strukturerschließung weitestgehend zu automatisieren und so mit begrenztem Aufwand eine beachtliche Datenmenge in eine Datenbank zu überführen.

Die ersten Vorläufer des Hoppenstedt-Aktienführers wurden im Jahr 1870 unter dem Namen „Saling's Börsenpapiere“ publiziert. Diese erschienen regelmäßig in verschiede-

[1] Siehe auch: Ausblick – Forschungsdaten an der UB Mannheim.



A

EDUARD AHLBORN AKTIENGESELLSCHAFT

Sitz: 32 Hildesheim, Lüntzelstraße 22, Postfach 530

Fernruf: Sa.-Nr. 8 32 71-75

Fernschreiber: 09 2763

Vorstand:
Ernst Morsch, Hildesheim, Vors.;
Dr. phil. Karl Bechtold, Hildesheim

Aufsichtsrat:
Ernst Hoeltje, Hannover, Vors.;
Dr. Werner Anders, Hannover, stellv. Vors.;
Justus Mundt, Freudenberg-Siegen;
Professor Dr.-Ing. Eduard Pestel, Hannover;
Achim Seibert, Bernried;
Bernd Wagner, Hildesheim;
Arbeitnehmervertreter:
Franz Atenhan, Hildesheim;
Theodor Mannes, Borsum;
Walter Mundry, Hildesheim

Gründung: 1927

Stückelung:
3 000 Inh.-St.-Akt. zu je DM 1 000.-

Großaktionär: Familienbesitz (ca. 60 %);
Rest Streubesitz

Aktienkurse (Hannover):
Notierung seit 9. 2. 1955

ultimo	1955	130	% +)
"	1956	128	%
"	1957	136	%
"	1958	185	%
"	1959	355	%
"	1960	570	%
"	1961	370	%
"	1962	301	%
30. Sept.	1963	341	%

+) ab Tag der Notierung Kurs für DM-Nennwert

Dividenden auf Stammaktien:
II/1948/49-1958: insgesamt 59 %
1959: 12 % (Div. Sch. Nr. 6)
1960: 13 % (Div. Sch. Nr. 7)
1961 u. 1962: je 12 % + 2 % Bonus (Div. Sch. Nr. 8 u. 9)

Abbildung 2.: Ausschnitt eines Unternehmensprofils (Screenshot)

nen Teilen für die verschiedenen deutschen Börsenplätze und auch zu Auslandsbörsen. Im Unterschied zu den Nachkriegs-Aktienführern ist die Information hier aber weniger strukturiert, eher in Fließtext enthalten. Da dies für automatisierte Verfahren eine Herausforderung darstellt, wurde im Aktienführer-Projekt hier zunächst auf die Erstellung einer Datenbank verzichtet. Aber alle Bände wurden digitalisiert und mit OCR bearbeitet, so dass zumindest eine Volltext-Suche verfügbar ist, die eine schnelle Durchsicht der einzelnen Bände ermöglicht.

Die Saling's Börsenpapiere umfassen zeitlich sehr prägende Epochen der deutschen und europäischen Geschichte, was sie zu herausragenden Quellen der Wirtschafts- und Sozialgeschichte macht. Man denke hier an die Industrialisierung im Kaiserreich, die Auswirkungen des Ersten Weltkriegs und die Weltwirtschaftskrise während der Weimarer Republik.

3. Vom Papier zur Datenbank – Erschließungsverfahren für Forschungsdaten am Beispiel des Aktienführer-Datenarchivs

Auf dem Weg vom gedruckten Werk zur Datenbank müssen eine Reihe von Schritten bewältigt werden. Zuerst müssen die Rechte in Zusammenarbeit mit dem Rechteinhaber geklärt werden. Dann sind die gedruckten Werke fachgerecht und mit hoher Qualität zu scannen, so dass dann Bilddaten des Originals vorliegen. Für die Erfassung der in den Bilddaten enthaltenen Informationen gibt es nun zwei Wege – händisch oder automatisch. Bei der händischen Datenerfassung übertragen Menschen die Informationen aus den Bilddaten in ein geeignetes elektronisches Format. Bei der automatischen Datenerfas-

sung übernimmt diese Aufgabe der Computer, wobei die entsprechenden Software-Tools dazu entwickelt werden müssen. Beide Wege wurden im Aktienführer-Projekt beschritten. In der ersten Projektphase entschied man sich für eine händische Datenerfassung, in der zweiten Projektphase dann für ein automatisiertes Vorgehen. Diese Abläufe sollen nun anhand des Aktienführer-Projekts vorgestellt werden.

Rechte am Aktienführer des Hoppenstedt-Verlags ist die Firma Bisnode Inhaberin der Aktienführer-Urheberrechte. Die UB Mannheim erwarb von Bisnode das Recht, alle Bände des Aktienführers sowie seiner Vorgänger zu digitalisieren, im Volltext strukturiert zu erfassen und wissenschaftlichen Einrichtungen kostenfrei zur Verfügung zu stellen. Die Einschränkung auf eine nicht-kommerzielle Nutzung war Bisnode sehr wichtig, um eigene Geschäftsmodelle nicht zu gefährden. Damit kann das Aktienführer-Datenarchiv nur wissenschaftlichen Einrichtungen und Einzelpersonen mit nachgewiesenem wissenschaftlichem Interesse verfügbar gemacht werden. Konkret wird der Zugang über eine Nationallizenz in Verwaltung der ZBW (Deutsche Zentralbibliothek für Wirtschaftswissenschaften) organisiert.[2]

Als erster Schritt zum Aktienführer-Datenarchiv waren alle verfügbaren Bände des Aktienführers und seiner Vorgängerpublikationen zu scannen. Da nicht alle Bände an der UB Mannheim vorhanden waren, wurde die Unterstützung durch Leihgabe und Geschenke anderer Bibliotheken benötigt.[3] Die Scanarbeiten wurden intern durch das UB-Digitalisierungsteam bewältigt.

Für das Aktienführer-Datenarchiv mit den Jahren 1870 bis 2018 wurden insgesamt 173 Bände mit zusammen 237.000 Seiten gescannt. Da für die Datenbank nur die Bände von 1956 bis 1999 genutzt wurden, entfallen auf diesen Teil 44 Bände mit etwa 50.000 Seiten.

Für die Jahre 2000 bis 2018 wurden die CD-ROM Ausgaben des Aktienführers verwendet. Hier war es nötig, eine Reihe von speziellen Software-Tools zu schreiben, um die verschlüsselten Daten zugänglich zu machen und strukturiert ins Aktienführer-Datenarchiv zu übertragen.

Händische Datenerfassung in Projektphase I

In der ersten Projektphase (2013–2015) bestand das Ziel darin, die Aktienführer-Bände für die Jahre 1976 bis 1999 zu digitalisieren und in eine Datenbank zu überführen. Wie schon gesagt, entschied man sich an dieser Stelle für eine händische Datenerfassung mittels „Double Keying“, da automatisierte Methoden zu dieser Zeit noch nicht die gewünschte Zeichen- und Strukturqualität erreichen konnten.

Beim „Double Keying“ werden die betreffenden Inhalte mehrfach händisch abgeschrieben, mit Vergleich und Korrektur der Ergebnisse. Die UB Mannheim beauftragte einen indischen Dienstleister mit dieser Aufgabe. Grundlage der Zusammenarbeit war ein detaillierter Erfassungsleitfaden, der für jede Datenkategorie die Regeln festhielt, wie die entsprechenden Daten zu erfassen sind. Schließlich lieferte der Dienstleister die struktu-

[2] URL Nationallizenz des Aktienführer-Datenarchivs: `https://www.nationallizenzen.de/angebote/nlproduct.2014-03-03.9100427542?`

[3] Ein herzlicher Dank geht hier an die USB Köln, die UB Bochum, die Bibliothek der HU Berlin, die UB Heidelberg, die FH-Bibliothek Würzburg, die Lippische Landesbibliothek Detmold, die ZBW, die SUB Göttingen, die Bibliothek des Deutschen Bundestags, die UB Greifswald sowie die StB Koblenz.

rierten Daten im XML-Format an die UB Mannheim. Nach einer ersten Testphase war die gelieferte Datenqualität überwiegend gut, wobei eine ständige Qualitätskontrolle von Seiten der UB notwendig blieb. Im Ergebnis lagen die Daten der Aktienführer-Jahrgänge 1976 bis 1999 in einem strukturierten Format vor.

Ein wichtiger Zwischenschritt an dieser Stelle war die Zusammenführung der Unternehmensprofile der verschiedenen Jahrgänge trotz unterschiedlicher Schreibweisen oder Umbenennungen der Unternehmen, um einfache und konsistente Zeitreihenanalysen zu ermöglichen.

Abschließend wurde aus den Anforderungen der Forscherinnen und Forscher aus dem Bereichen Finance sowie Accounting & Taxation eine sinnvolle Datenbankstruktur und Weboberfläche entwickelt, um eine komfortable Nutzung der Forschungsdaten zu ermöglichen. Im Zentrum stand dabei die Möglichkeit, individuell zusammengestellte Datenpakete in einem Standardformat (hier CSV) zur direkten Weiterverarbeitung herunterladen zu können.

Im Ergebnis der ersten Projektphase standen die Daten des Aktienführers für die Jahre 1976 bis 1999 strukturiert zur Verfügung. Es wurde aber schnell klar, dass diese zeitliche Abdeckung für die Forschung nicht ausreichend ist. Dank weiterer DFG-Förderung konnte dem in einer zweiten Projektphase von 2017 bis 2019 abgeholfen werden.

Automatisierte Datenerfassung in Projektphase II

Auch bei der automatisierten Datenerfassung bildeten die beim Scannen erzeugten Bilddaten den Ausgangspunkt der Arbeit. Im Unterschied zum händischen Vorgehen war jetzt aber eine Reihe von Zwischenschritten zu absolvieren, um die Bilddaten in Text und schließlich in eine strukturierte Datenbank umzuwandeln. Eine Übersicht dazu gibt die folgende Abbildung 3.

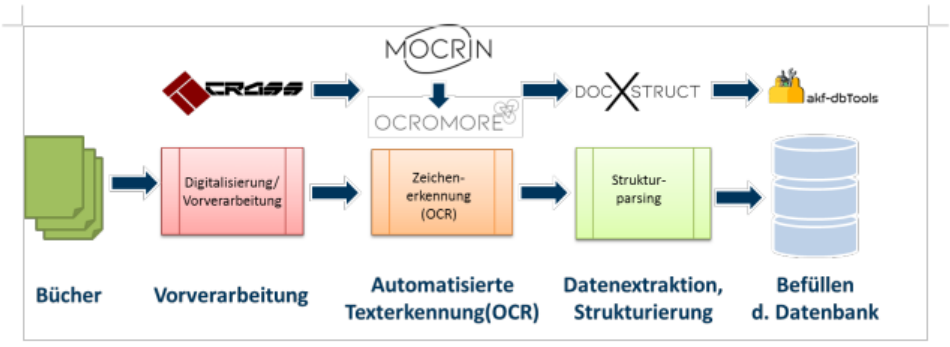


Abbildung 3.: Arbeitsschritte zur automatisierten Datenerfassung (eigene Darstellung)

Die in Abb. 3 dargestellten vier Arbeitsschritte sollen nun im Einzelnen dargestellt werden. Für alle Arbeitsschritte hat das Projektteam der UB Mannheim spezielle Software-Tools geschrieben, die frei verfügbar auf GitHub sind und als Open Source nachgenutzt werden können.[4]

[4] Die Aktienführer-Datenarchiv-Tools auf GitHub: `https://github.com/UB-Mannheim/`Aktienfuehrer-Datenarchiv-Tools

Schritt 1 – Vorverarbeitung

Mittels des Tools „crass“ wurden die gescannten Bilddaten für die folgende Texterkennung vorbereitet. Dazu wurde die 2-Spaltenstruktur der Aktienführer-Unternehmensprofile aufgelöst und in eine leichter zu verarbeitende 1-Spaltenstruktur überführt. Hinzu kommen das Zusammenfassen einzelner Textsegmente und die Korrektur von Verzerrung, die beim Scannen erfolgt sein können.

Schritt 2 – Automatisierte Texterkennung (OCR)

Im zweiten Schritt geht es darum, in den vorliegenden Bilddaten den enthaltenen Text zu erkennen. Zu diesem Zweck wurden im Aktienführer-Projekt drei verschiedene, populäre OCR-Programme parallel genutzt: Tesseract, Ocropus und ABBYY FineReader. Das heißt, jede Bilddatei wurde mehrfach mit OCR bearbeitet mit dem Ziel, die Stärken aller drei OCR-Programme zu kombinieren. Dies bedeutet aber auch, dass die entstandenen drei OCR-Ergebnisse in einem weiteren Schritt wieder zu einem gemeinsamen Ergebnis zusammengefügt werden müssen, welches möglichst die vorhandenen Erkennungsfehler minimiert.

Um diesen Prozess abzubilden, wurden zwei spezielle Tools geschrieben: Mocrin und Ocromore.

Mocrin ermöglicht, die drei OCR-Programme Tesseract, Ocropus und FineReader über ein einziges Interface zu steuern und die Ergebnisse im hOCR-Dateiformat in einer einheitlichen Ordnerstruktur abzulegen.[5] Dabei sind die Erkennungsergebnisse mit Konfidenzen versehen. Diese Konfidenzen geben an, mit welcher Sicherheit ein bestimmtes Zeichen vom OCR-Programm erkannt wurde.

Ocromore übernimmt dann die Zusammenführung zu einem optimierten Endergebnis auf Basis der von den OCR-Programmen gelieferten Konfidenzen. Ein Beispiel für diesen Prozess ist in Abbildung 4 dargestellt.

Darin wird deutlich, wie Ocromore die Ergebnisse der drei OCR-Programme (R1, R2, R3) vergleicht und zusammenführt. Am Beispiel des dritten Buchstabens aus „Eduard“ sieht man, dass R1 ein „u“ mit einer Konfidenz von 99 Prozent erkannt hat, R2 erkennt ein „o“ mit 60 Prozent Konfidenz und R3 wieder ein „u“ mit 90 Prozent Konfidenz. Zusammengefasst gibt das für „u“ einen Konfidenzwert von 189 Punkten und für „o“ nur 60 Punkten, weswegen Ocromore das „u“ an dieser Stelle auswählt. Auf diese Art und Weise gelingt es Ocromore, die Erkennungsergebnisse zu vergleichen und zusammenzuführen.

Dieses Vorgehen ermöglicht eine deutliche Erhöhung der Erkennungsgenauigkeit, wie in Abbildung 5 dargestellt.

Dabei wird für den Aktienführer eine Erkennungsgenauigkeit von 99,6 Prozent erreicht, was einer Fehlerreduktion von ca. 33 % im Vergleich zum besten der einzelnen OCR-Ergebnissen entspricht. Ebenfalls sieht man eine Erhöhung der Erkennungsgenauigkeit für den englischen Standardkorpus UNLV (University of Nevada Las Vegas Standardized Test Set).[6]

[5] Das hOCR-Format ist ein offener Standard zur Darstellung von OCR-Ergebnissen, vgl. Breuel 2007 und `http://kba.cloud/hocr-spec/1.2/`.

[6] Zu Ocromore vgl. auch Kamlah /Stegmüller 2018.

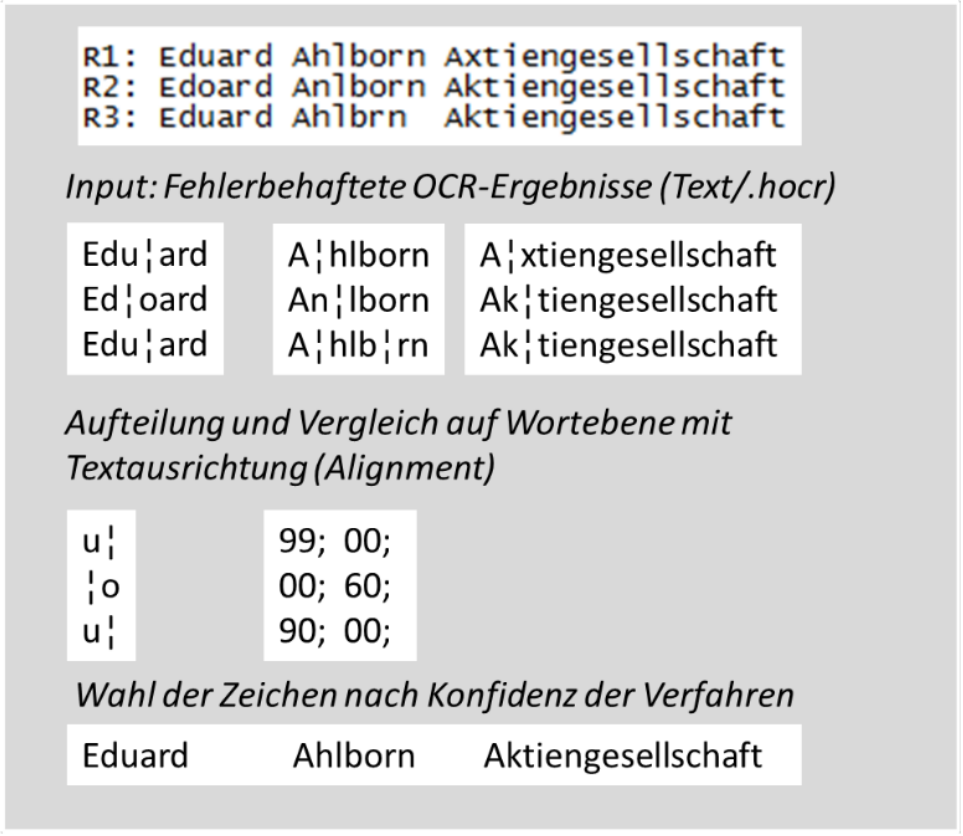


Abbildung 4.: Evaluierung und Zusammenführung der OCR-Ergebnisse durch Ocromore (eigene Darstellung)

Im Ergebnis von Schritt zwei wurde der Text automatisiert erkannt und ist nun im hOCR-Format zur Weiterverarbeitung parat.

Schritt 3 – Datenextraktion und Strukturierung

Mittels des Tools Docxstruct werden die hOCR-Dateien, die das Ergebnis der automatisierten Texterkennung sind, analysiert, kategorisiert und segmentiert und in das JSON-Format überführt. An dieser Stelle wird erkannt, in welche Datenbankbereiche der erkannte Text passt, zu welcher Kategorie die Information also gehört. Ein Beispiel dafür gibt die folgende Abbildung 6.

Man sieht, wie der erkannte Text (origpost) ausgelesen und in einzelne Kategorien (numID, city, street, additional info) aufgelöst wird, die sich dann entsprechend in der Datenbank wiederfinden. Damit gelingt es in Schritt 3, den erkannten Text umfassend zu strukturieren – ihn also den passenden Bereichen der Datenbank zuzuordnen.

Schritt 4 – Befüllen der Datenbank

Nachdem die Daten ausgelesen und strukturiert wurden, können sie in einem letzten Schritt in die in Projektphase I erstellte Datenbank geschrieben werden. Dazu wurde das

OCR-Engine	Aktienführer (AKF)	UNLV
ABBYY	99,35 %	98,46 %
OCRopus (default en-model)	-	92,49 %
OCRopus (trained)	98,76 %	-
Tesseract	99,00 %	98,23 %
ocromore (MSA)	**99,60 %**	**98,65 %**

Abbildung 5.: Erhöhte Erkennungsgenauigkeit durch Ocromore (eigene Darstellung)

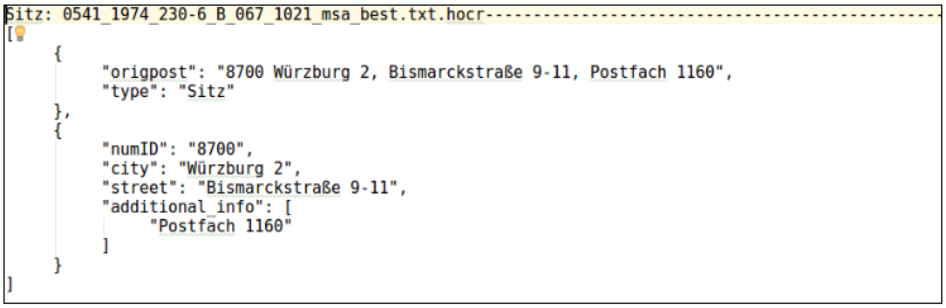


Abbildung 6.: Strukturerkennung mit Docxstruct (eigene Darstellung)

Tool akf-db-Tools geschrieben, eine Sammlung kleinerer Skripte für die Befüllung der Datenbank. Dabei erfolgen auch noch Schritte zur Normalisierung und Deduplizierung der Daten, um die Qualität der Forschungsdaten zu erhöhen.

Im Ergebnis der automatisierten Datenerfassung wurde das Aktienführer-Datenarchiv um 20 Jahrgänge (1956–1975) ergänzt. Zusammen mit den Bänden aus Projektphase I (1976–1999) und den Daten aus den CD-ROMs (2000–2018) ergibt sich eine zusammenhängende Abdeckung über 62 Jahre.[7]

Die Darstellung der automatisierten Datenerfassung zeigt, dass dafür eine Reihe von Schritten und Softwareentwicklungen nötig war. Es wird deutlich, dass für ein solches spezialisiertes Projekt im Moment keine „out-of-the-box“ Lösung verfügbar ist. Nur mit genügend Ressourcen in Form von Zeit und Know-how konnte dieser Weg beschritten werden. Als Gewinn neben den erschlossenen Forschungsdaten erscheinen aber auch die frei nachnutzbaren Software-Tools, die bei zukünftigen, ähnlichen Projekten deutliche Einsparungen versprechen.

Mit Abschluss der zweiten Projektphase steht nun das vollständige Aktienführer-Datenarchiv elektronisch zur Nutzung bereit.[8] Das Aktienführer-Datenarchiv bietet dabei drei Nutzungsmöglichkeiten:

[7] Aus lizenzrechtlichen Gründen können die Jahrgänge 2017 und 2018 allerdings erst 2020 bzw. 2021 zugänglich gemacht werden.

[8] URL des Aktienführer-Datenarchivs: `https://digi.bib.uni-mannheim.de/aktienführer/data/index.php`

- **Schnellsuche** – hier kann die Datenbank (Jahre 1956 – 2018) über die Datenfelder Firmenname, Personen und Wertpapierkennnummer (WKN) durchsucht werden. Es werden dann Treffer für gefundene Unternehmensprofile, Beteiligungen, Personen und WKN angezeigt. Der Zweck der Schnellsuche ist es, einen schnellen Überblick über die verfügbaren Daten zu gewinnen.

- **Datenexport** – hier können große Datenpakete für empirische Analysen zusammengestellt und heruntergeladen werden. Der Nutzer kann dabei die benötigten Unternehmen, Jahre und Datenkategorien wählen und das Ergebnis im CSV-Format herunterladen. Das CSV-Format ist ein Standard im Bereich der quantitativen Datenanalyse und kann direkt in Auswertungsprogramme wie Stata, R oder SPSS aber auch Microsoft Excel oder LibreOffice Calc geladen werden.

- **Scans** – hier stehen die Volltext-Digitalisate der Print-Bände des Hoppenstedt-Aktienführers und seiner Vorgängerpublikationen (Jahre 1870 – 1999) zur Nutzung bereit. Alle Bände sind mit Inhaltsverzeichnissen versehen und können im Volltext durchsucht und heruntergeladen werden. Somit ist eine komfortable und effiziente Nutzung möglich. Falls bei der Nutzung der Datenbank Zweifel an den präsentierten Daten bestehen, können alle Angaben mithilfe der Scans verifiziert werden.

In Abbildung 7 sieht man am Beispiel des Unternehmens Daimler-Benz die Oberfläche zum Export der Aktienkurse aus den Jahren 1970 bis 1980.

Nach dem Export stehen die Daten des Aktienführer-Datenarchivs unmittelbar zur Auswertung bereit. Ein Nutzungsbeispiel könnte die Frage sein, wie sich die Ölkrise der 70er Jahre auf die Firma Daimler-Benz AG ausgewirkt hat. Abbildung 8 präsentiert eine Darstellung der aus dem Aktienführer gewonnenen Daten, die zeigt, dass zwar der Aktienkurs zeitweilig schwankte, Umsatz und Jahresüberschuss aber keine gravierenden negativen Entwicklungen aufwiesen. Größere Datensets dieser Art können von der empirischen Forschung zum Testen von Hypothesen und Theorien verwendet werden.

Das Aktienführer-Datenarchiv hat durch die Ergänzung der Daten der zweiten Projektphase erheblich an Wert für die Forschung gewonnen. Schon bisher zeigten sehr gute Nutzungsdaten die Bedeutung des Angebotes an. So haben sich bisher 142 Einrichtungen für die Nutzung über die Nationallizenz angemeldet. Außerdem gibt es auch noch mehr als 380 persönliche Accounts für Personen, die ein berechtigtes Interesse nachweisen konnten und ohne institutionelle Zugriffsmöglichkeit sind. Die Webseite erhielt im Jahr 2018 etwa 12.500 Besuche. Es ist zu erwarten, dass das deutlich erweiterte Datenangebot zu noch mehr Nutzung führen wird.

Im Endergebnis stehen nun mit dem Aktienführer-Datenarchiv hochrelevante Forschungsdaten für die Wirtschaftswissenschaften bereit. Eine bisher nur gedruckt vorliegende Publikation wurde durch die Anstrengungen der UB Mannheim in Forschungsdaten transformiert, die nach den heutigen Standards der Forschung genutzt werden können. Es steht zu erwarten, dass hieraus viele interessante Publikationen erwachsen.

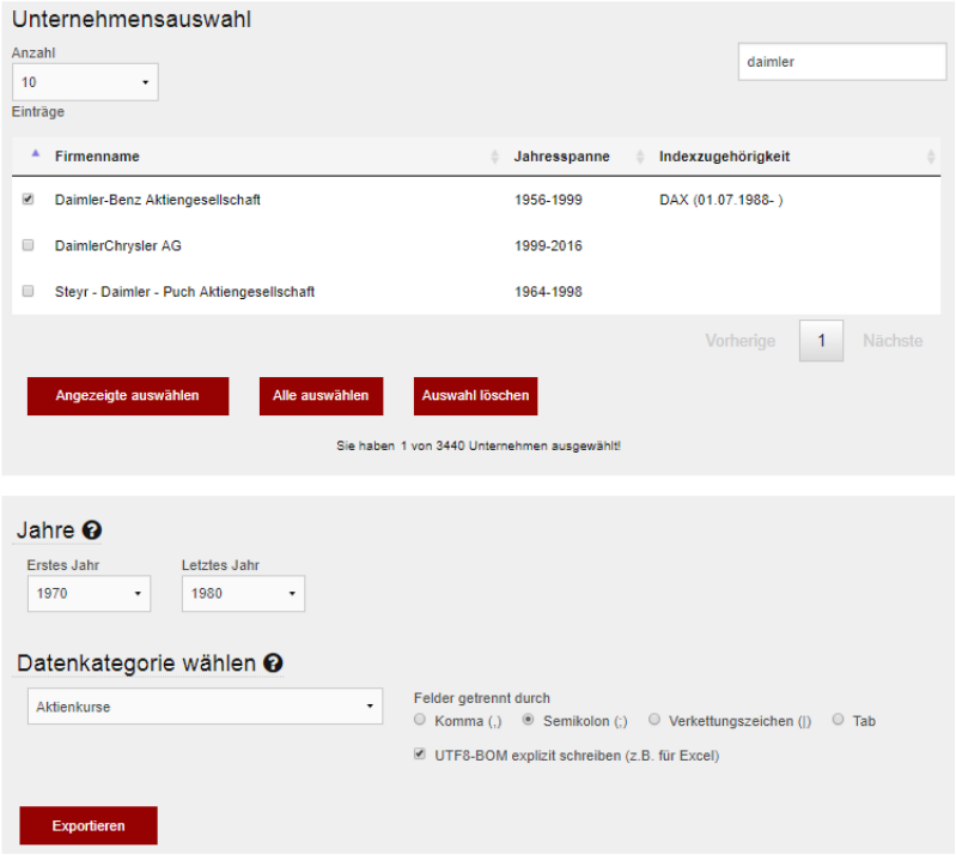


Abbildung 7.: Datenbank-Export für Daimler-Benz (eigene Darstellung)

4. Ausblick – Forschungsdaten an der UB Mannheim

Mit dem Aktienführer-Datenarchiv hat die UB Mannheim einen ersten, richtungsweisenden Schritt im Bereich der wirtschaftswissenschaftlichen Forschungsdaten getan, dem noch weitere folgen werden.

Aufbauend auf dem Aktienführer-Datenarchiv sollen weitere wirtschaftswissenschaftliche Forschungsdaten aus dem Hause Hoppenstedt / Bisnode erschlossen und bereitgestellt werden. Die UB Mannheim hat von dem Anbieter die Digitalisierungsrechte an einer Reihe weiterer fachlich relevanter Publikationen erworben.[9] Auch diese sollen nach dem Vorbild des Aktienführer-Datenarchivs digitalisiert und erschlossen werden. Man kann erwarten, dass hier ein Portfolio hochrelevanter Forschungsdaten zur Entwicklung der deutschen Wirtschaft entsteht.

Diese Forschungsdaten sind der Grundbaustein für das Forschungsdatenzentrum (FDZ) an der Universität Mannheim. Im Rahmen des FDZ werden auch andere Forschungsdaten angeboten, wie zum Beispiel der „Deutsche Reichsanzeiger und Preußische Staatsanzeiger“.

[9] Zu diesen Publikationen gehören: „Leitende Männer und Frauen der Wirtschaft“, „Großunternehmen“, „Mittelständische Unternehmen“, „Handbuch der deutschen Aktiengesellschaften : das Spezial-Archiv der Deutschen Wirtschaft“.

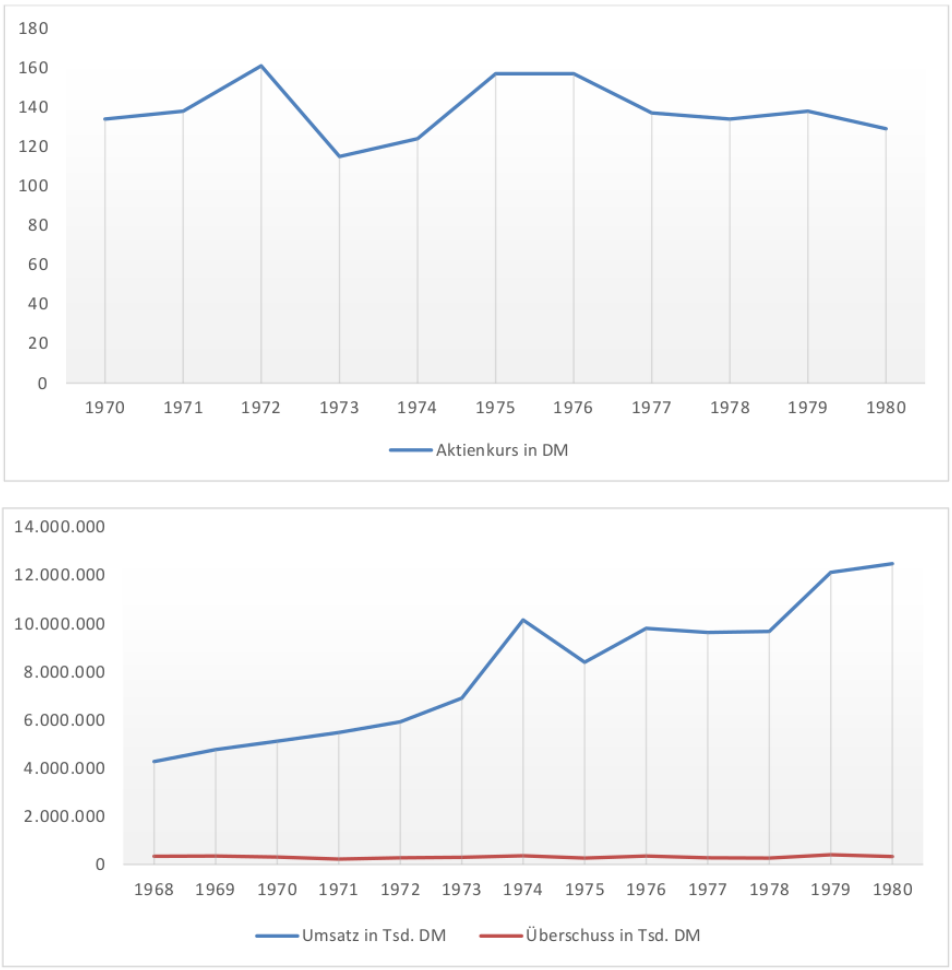


Abbildung 8.: Aktienkurs, Umsatz und Jahresüberschuss der Daimler-Benz AG in den 70er Jahren (eigene Darstellung)

Um die Kompetenz für OCR-Erschließung von vorhandenen Bilddigitalisaten noch breiter an Bibliotheken und anderen Kultureinrichtungen des Landes zu verankern, wird die UB Mannheim gemeinsam mit der Universitätsbibliothek Tübingen zudem in den kommenden beiden Jahren Tools und Beratungsservices bereitstellen, die den Einsatz automatisierter Erschließungsverfahren auch ohne einschlägiges Expertenwissen ermöglichen sollen. Das Projekt OCR-BW schließt an das nationale, DFG-geförderte Projekt OCR-D an, bei dem die UB Mannheim ebenfalls beteiligt ist, und wird durch das Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg gefördert.

Ab Mitte 2019 wird die Universitätsbibliothek Mannheim gemeinsam mit Wissenschaftlerinnen und Wissenschaftlern der Universität Mannheim und des Leibniz-Zentrums für Europäische Wirtschaftsforschung (ZEW) ein Kompetenzzentrum für Datenverfügbarkeit und Datenanalyse in den Wirtschaftswissenschaften für Baden-Württemberg aufbauen, das als eines von vier großen Science Data Centers durch das Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg gefördert wird. Das Business and Economic Research Data Center (BERD-Center BW) wird den Zugang zu vor-

handenen wirtschaftswissenschaftlichen Datenbeständen verbessern und neue, unstrukturierte Datenquellen (Big Data) erschließen. Damit kann auch ein wichtiger Beitrag für eine künftige Nationale Forschungsdateninfrastruktur (NFDI) und eine European Open Science Cloud geleistet werden, deren Ziel eine umfassende, verlässliche, gut zugängliche und möglichst offene, vernetzte Infrastruktur für Forschungsdaten ist.

Literaturverzeichnis

[1] Breuel, Thomas M. (2007): „The hOCR Microformat for OCR Workflow and Results“. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2:1063–67. Online:https://doi.org/10.1109/ICDAR.2007.4377078 [02.05.2019].

[2] Kamlah, Jan und Stegmüller, Johannes (2018): Ocromore – Combining multiple OCR-engine results to improve character recognition accuracy. Poster Bibliotheca Baltica Symposium 2018 (BB2018), Rostock, 3.-5.October 2018. Online: http://doi.org/10.5281/zenodo.1493860 [02.05.2019].

[3] item Zumstein, Philipp (2016): Die Bibliothek als Daten-Jongleur. Vortrag beim Bibcast 2016. Online:
http://bibcast.openbiblio.eu/die-bibliothek-als-daten-jongleur-services-fuer-datenzentrierte-forschung/ [02.05.2019].