OCROMORE

Combining multiple OCR-engine results to improve character recognition accuracy

A BRIEF INTRODUCTION

Ocromore is a toolset to increase the character recognition accuracy by combining multiple ocr results. The character accuracy is especially **crucial for automatic structuring and research purposes**.

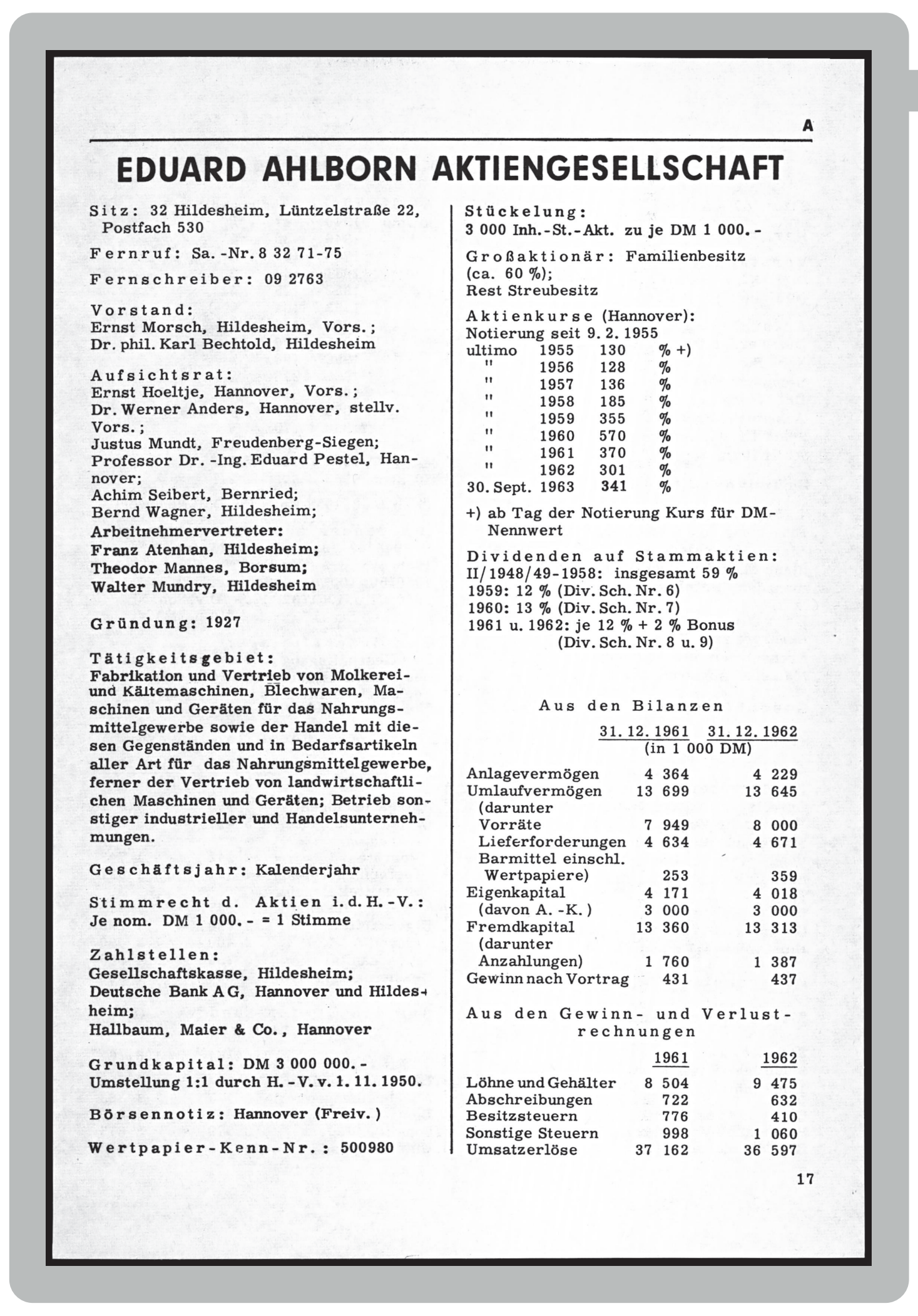
It was developed for the **DFG-Project Aktienführer-Datenarchiv**.
The Aktienführer is a reference work published annually between 1956-1999 as print book comprising data for companies listed at stock exchanges in Germany.

One of the goal of this project is to process this data and store it in a structured manner in a **database**. To optimize the OCR workflow and the result, several tool where created and combined to a **toolchain**.

In the **preprocessing** step, each company record get deskewed, splitted by single segments and merged into one image with one coloumn textblock with „**crass**".

The produced images are then used to create multiple **OCR** results. To simplify the communication and guarantee a unified project structure a multiple OCR-engine interface „**akf-mocrin**" was designed.

The **postprocessing** step of „ocromore" is described in detail the following section.
The content of the combined result is further structered with „**hocr-parser**".



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

Tätigkeitsgebiet:
Fabrikation und Vertrieb von Molkerei- und Kältemaschinen, Blechwaren, Maschinen und Geräten für das Nahrungsmittelgewerbe sowie der Handel mit diesen Gegenständen und in Bedarfsartikeln aller Art für das Nahrungsmittelgewerbe, ferner der Vertrieb von landwirtschaftlichen Maschinen und Geräten; Betrieb sonstiger industrieller und Handelsunternehmungen.

Geschäftsjahr: Kalenderjahr

Stimmrecht d. Aktien i.d.H.-V.:
Je nom. DM 1 000.- = 1 Stimme

Zahlstellen:
Gesellschaftskasse, Hildesheim;
Deutsche Bank AG, Hannover und Hildesheim;
Hallbaum, Maier & Co., Hannover

Grundkapital: DM 3 000 000.-
Umstellung 1:1 durch H.-V. v. 1. 11. 1950.

Börsennotiz: Hannover (Freiv.)

Wertpapier-Kenn-Nr.: 500980

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)

Aus den Bilanzen

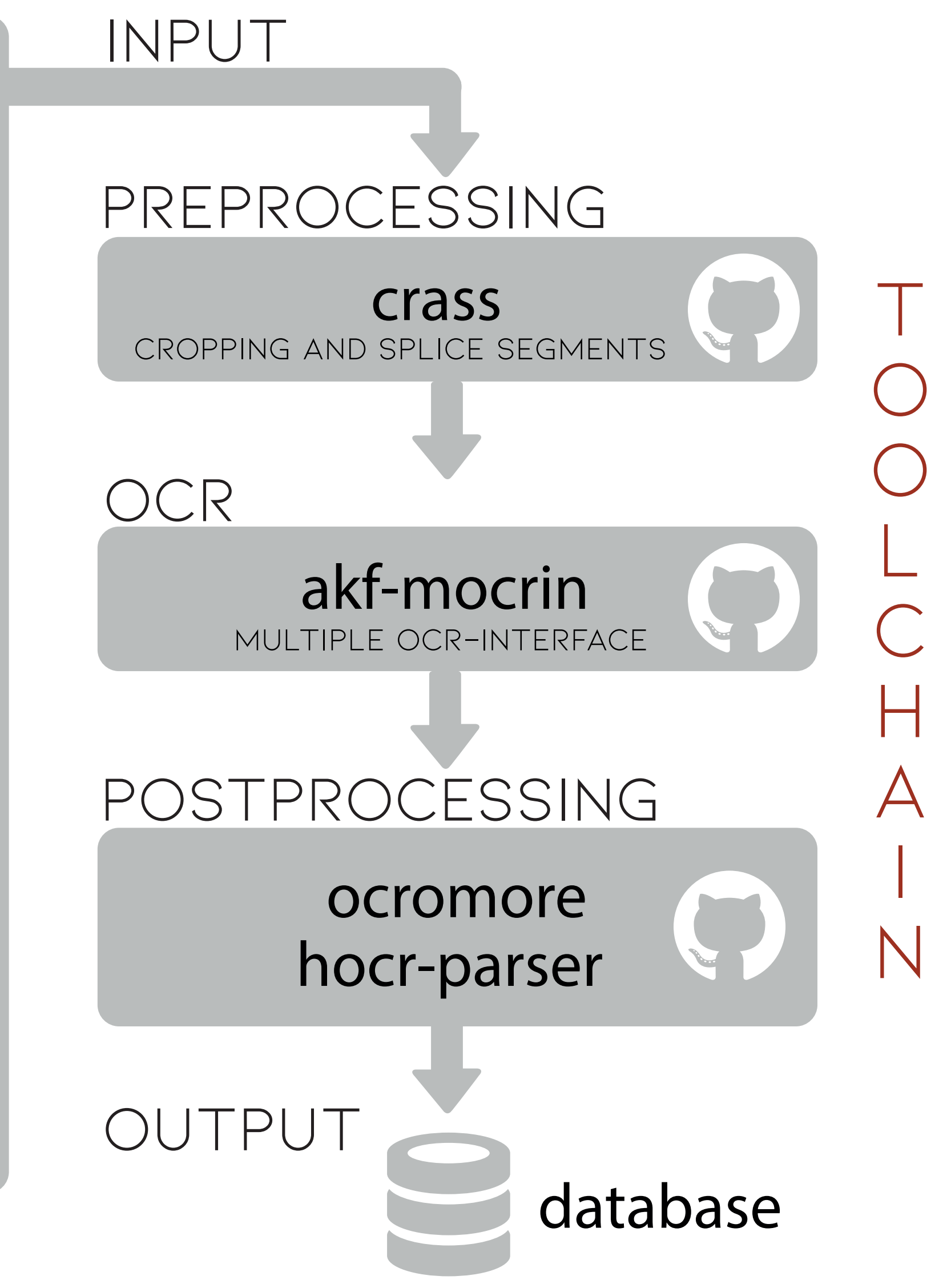
	31. 12. 1961	31. 12. 1962
	(in 1 000 DM)	
Anlagevermögen	4 364	4 229
Umlaufvermögen	13 699	13 645
(darunter		
Vorräte	7 949	8 000
Lieferforderungen	4 634	4 671
Barmittel einschl.		
Wertpapiere)	253	359
Eigenkapital	4 171	4 018
(davon A.-K.)	3 000	3 000
Fremdkapital	13 360	13 313
(darunter		
Anzahlungen)	1 760	1 387
Gewinn nach Vortrag	431	437

Aus den Gewinn- und Verlustrechnungen

	1961	1962
Löhne und Gehälter	8 504	9 475
Abschreibungen	722	632
Besitzsteuern	776	410
Sonstige Steuern	998	1 060
Umsatzerlöse	37 162	36 597

17

Aktienführer company record



INPUT

PREPROCESSING

crass
CROPPING AND SPLICE SEGMENTS

OCR

akf-mocrin
MULTIPLE OCR-INTERFACE

POSTPROCESSING

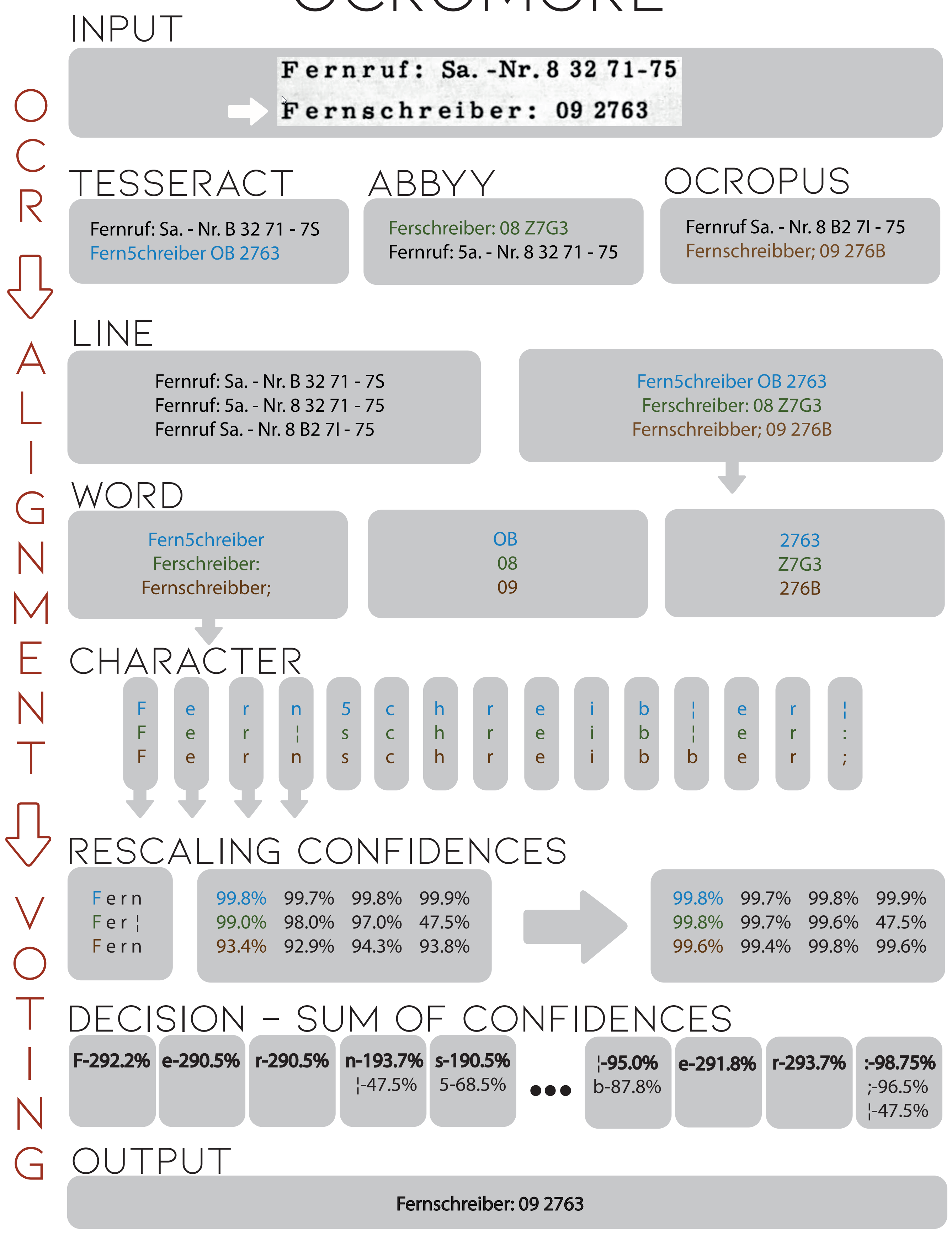
ocromore
hocr-parser

OUTPUT

database

TOOLCHAIN

OCROMORE



INPUT

Fernruf: Sa. -Nr. 8 32 71-75
Fernschreiber: 09 2763

OCR

TESSERACT
Fernruf: Sa. - Nr. B 32 71 - 7S
Fern5chreiber OB 2763

ABBYY
Ferschreiber: 08 Z7G3
Fernruf: 5a. - Nr. 8 32 71 - 75

OCROPUS
Fernruf Sa. - Nr. 8 B2 7l - 75
Fernschreibber; 09 276B

ALIGNMENT

LINE
Fernruf: Sa. - Nr. B 32 71 - 7S
Fernruf: 5a. - Nr. 8 32 71 - 75
Fernruf Sa. - Nr. 8 B2 7l - 75

Fern5chreiber OB 2763
Ferschreiber: 08 Z7G3
Fernschreibber; 09 276B

WORD
Fern5chreiber
Ferschreiber:
Fernschreibber;

OB
08
09

2763
Z7G3
276B

CHARACTER

F	e	r	n	5	c	h	r	e	i	b	¦	e	r	¦
F	e	r	¦	s	c	h	r	e	i	b	¦	e	r	:
F	e	r	n	s	c	h	r	e	i	b	b	e	r	;

VOTING

RESCALING CONFIDENCES

Fern	99.8%	99.7%	99.8%	99.9%
Fer¦	99.0%	98.0%	97.0%	47.5%
Fern	93.4%	92.9%	94.3%	93.8%

99.8%	99.7%	99.8%	99.9%
99.8%	99.7%	99.6%	47.5%
99.6%	99.4%	99.8%	99.6%

DECISION – SUM OF CONFIDENCES

F-292.2%	**e-290.5%**	**r-290.5%**	**n-193.7%** ¦-47.5%	**s-190.5%** 5-68.5%	•••	**¦-95.0%** b-87.8%	**e-291.8%**	**r-293.7%**	**:-98.75%** ;-96.5% ¦-47.5%

OUTPUT

Fernschreiber: 09 2763

WORKFLOW OR THREE FOR ONE

From the images, like the example left, initally different **OCR**-results are generated. This project use the OCR-engines, **Tesseract**, **Ocropus** and **Abbyy Finereader**.

It is necessary to produce results with word bounding boxes for the **alignment** and character confidence for the **voting** process. The results differ not only in single recognized characters but also in the arrangment of the lines. This leads to the first section, the **alignment**.

The **alignment** splits into three subsection: **line**, **word** and **character**.
While the first two alignment steps are based on bounding boxes, the third takes a whole other turn. To match all lines and words to their pendants a mean value with some tolerance of horizontal and vertical bounding box values are calculated. Although the wordspaces are now mapped to one another, the content is not.

For the **character** level, a multi sequence alignment algorithm, called msa, was implemented.
Msa is frequently applied to find matching subsequences in biological sequences. Using msa with very long sequences and sequences containing more variety than a few base pairs like natural language (UTF-8) can be prohibitively computationally expensive and/or can lead to unmatchable combinations. For that reason, splitting into wordwise fragments is quite important. The algortihm was additionally extended, that not only the same character but also optically similar get aligned. When matching the subsequcenes gaps are created, which are filled with so-called **wildcards** „¦".

In the last section, the **voting** process, every matched character set is subjected to a decision process based on the confidence. Because each engine is weighted differently, rescaling the confidences in the first place is essential. It turned out that it is worth taking a close look, as some OCRs have special recognition strength. In the rescaling process letters, numbers, symbols and even special characters can be weighted differently.

The decision making process sums up all confidences and picks the highest one. If a wildcard is picked, it will be removed and the text left and right get concatenated.

The whole process can be declared as
character-confidencebased wordwise multi-sequence alignment method.

RESULTS – TRIED AND TESTED

The gained results are evaluated against groundtruth (GT) data with a standard tool (**ISRI Analytic Tools**).

Our evaluated datasets showed a character accuracy of 84,08% for Abbyy, 98,67% for Ocropus, 98,79% for Tesseract and 99,19 % for the combined result, which represents an **accuracy increase of 0,49%** compared to the best single result and an **error reduction of 33%**.

"Ocromore" and the other mentioned projects are written in Python and released under the patronage of the Mannheim University Library as open source projects in Github (see section „Open source projects.")
The Apache License permit everyone to reuse and adapt the code for their own use case.
Access to the database and more information about the DFG-Project Aktienführer-Datenarchiv:
https://digi.bib.uni-mannheim.de/aktienfuehrer/.

OCR-Engine	AKF-II	UNLV
Abbyy	84,08 %	88.85 %
Ocropus (default en-model)		87.33 %
Ocropus (trained)	98,67 %	
Tesseract	98,79 %	96.59 %
MSA	99,19 %	96.73 %

DFG-PROJEKT AKTIENFÜHRER II

JAN KAMLAH
JAN.KAMLAH@BIB.UNI-MANNHEIM.DE

JOHANNES STEGMÜLLER
JOHANNES.STEGMUELLER@BIB.UNI-MANNHEIM.DE

UNIVERSITÄT
MANNHEIM

UB
MANNHEIM

OPEN SOURCE PROJECTS

CRASS: CROPPING AND SPLICE SEGMENTS
HTTPS://GITHUB.COM/UB-MANNHEIM/CRASS

aKF-MOCRIN: MULTIPLE OCR-INTERFACE
HTTPS://GITHUB.COM/JKAMLAH/AKF-MOCRIN

OCROMORE:
HTTPS://GITHUB.COM/UB-MANNHEIM/OCROMORE

HOCR-PARSER:
HTTPS://GITHUB.COM/UB-MANNHEIM/HOCR-PARSER