


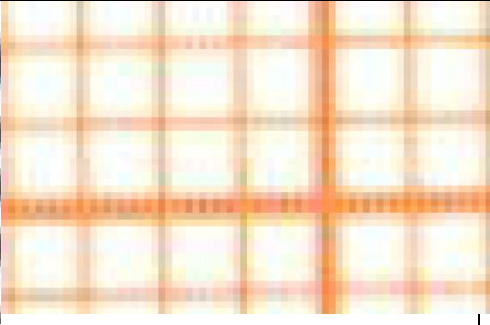




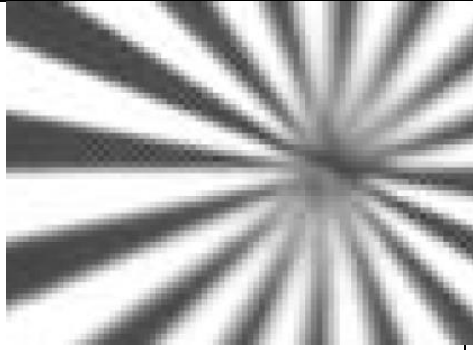
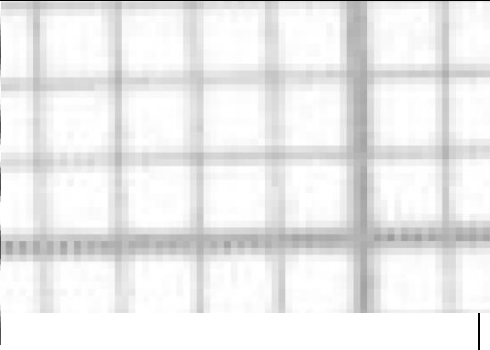



Dokumentation Best practice: Scannen mit verschiedenen Scannern + OCR-Umwandlung

Basis: Eine Doppelseite Frakturtext wurde mit unterschiedlichen Scannern („UB-Buchscanner“, „NDL-Kopierer“) und unterschiedlichen Parametern (Farbe, Auflösung „dpi“) gescannt. Im Anschluss daran wurde der Text mit dem Abbyy Fine Reader maschinenlesbar gemacht und die falsch erkannten Buchstaben gezählt („Anzahl der Fehler im Frakturtext“). Die Ergebnisse sind in der Tabelle dokumentiert. Um die Unterschiede besser zu erkennen, wurde außerdem das in der Archivierung als Standard bekannte „Sonnenbild“ und normales Millimeterpapier (Tipp von Klaus Wendel von www.archium.org) gescannt.

- Beim UB-Buchscanner lässt sich nur die Farbeinstellung ändern (Farbe; s/w; Graustufen). Die Auflösung ist nicht veränderbar (immer 300dpi). **Die Einstellung „Farbe“ und 300dpi ist beim UB-Buchscanner ideal**, da sie die geringste Fehlerquote aufweist, d.h. die wenigsten falsch erkannten Zeichen bei der OCR-Umwandlung.
 - Beim NDL-Kopierer lassen sich sowohl Farbeinstellung als auch Auflösung anpassen. Es wurden jeweils beide Parameter verändert. Die Auflösungen 200, 300, 400 und 600 dpi wurden jeweils mit Vollfarbe (256 Farben), s/w, Graustufen und der automatischen Farbeinstellung gescannt. Die geringste Fehlerquote und damit **das beste Ergebnis** wurde mit einer Einstellung **Vollfarbe, 400 dpi** erreicht.
- ➔ Die erste Wahl ist der NDL-Kopierer mit der Einstellung Vollfarbe, 400dpi, da Abbyy auf einer Doppelseite 2 Fehler weniger macht. Wenn mit dem UB-Scanner gescannt werden muss, dann mit der Einstellung „Farbe“ (und 300dpi, das sowieso nicht veränderbar ist).

In der Tabelle auf der letzten Seite sind die Fehler, die Abbyy beim Umwandeln gemacht hat, aufgelistet. Man sieht deutlich, dass man die Fehlerquote auf nahezu null reduzieren kann, wenn man als weiteren Schritt Abbyy auf die Fehler hin trainiert. D.h. die Buchstabenkombination „st“, die meistens als „si“ erkannt wurde wird als ein Zeichensatz trainiert, der eben „st“ bedeutet. Siehe dazu **02-Anleitung Abbyy Fine Reader OCR von Frakturtexten.doc**

UB-Buch-scanner	dpi	„Sonnenbild“	Millimeterpapier	Doppelseite in Fraktur	Anzahl der Fehler im Frakturtext
Farbe	300				7
s/w	300				17
Grau-stufen	300				9



NDL-Kopierer	dpi	„Sonnenbild“	Millimeterpapier	Doppelseite in Fraktur	Anzahl der Fehler im Frakturtext
Vollfarbe (256 Farben)	200				7
Auto-farbe	200				7
s/w	200				22
Grau-stufen	200				8
Vollfarbe (256 Farben)	300				8
Auto-farbe	300				8
s/w	300				9
Graustufen	300				8



Vollfarbe (256 Farben)	400				5
Auto- farbe	400				6
s/w	400				10
Grau- stufen	400				6
Vollfarbe (256 Farben)	600				8
Auto- farbe	600				7
s/w	600				8
Grau- stufen	600				7



Fehler bei der Umwandlung von einer Doppelseite in Abbyy:

		e → ¹ c	u →n	st →si	st →il	ss →st	st →ss	st →ff	b →d	m →n	l →!	s →j	s →f	b →h	t →f	i →l	c →:	e →o	e →h	L →T	i →t	h →y	lw →hn	e →r
UB-Buchscanner																								
Farbe	300			1							1	1	1	2	1									
s/w	300	1		6						3						1	1	1	1	1				
Graustufen	300	1		5	1	1			1															
NDL-Kopierer																								
Vollfarbe	200			5			1			1														
Autofarbe	200			5				1		1														
s/w	200	2	10	4		1		1		2			1											
Graustufen	200			5			1			1														
Vollfarbe	300			4			1	2		1														
Autofarbe	300			6																	2			
s/w	300			5				2		2														
Graustufen	300			5					1													2		
Vollfarbe	400			5																				
Autofarbe	400			5									1											
s/w	400			4			1	1		2												1	1	
Graustufen	400			4			1	1																
Vollfarbe	600			4			1	1		2														
Autofarbe	600			4				1		2														
s/w	600			5			1			2														
Graustufen	600			4			1			2														

*trainiert wurden nur st, und tz

** Zahlen wurden weggeschnitten

***jede Datei wurde neu trainiert

******Zeitaufwand** pro Doppelseite:
trainieren zwischen 15 und 30 Minuten
überprüfen zwischen 5 und 10 Minuten

¹ „→“ steht für „statt“