



Documentation Best practice: Scanning with different scanners + OCR conversion

Basis: A double page fracture text was scanned with different scanners ("UB book scanner","NDL copier") and different parameters (colour, resolution "dpi"). Afterwards, the text was made machine-readable with the Abbyy Fine Reader and the wrongly recognized letters were counted ("number of errors in the fracture text"). The results are documented in the table. In order to better recognize the differences, the "sun image" and normal graph paper (tip by Klaus Wendel from www.archium.org), which is known as the standard in archiving, were also scanned.

- With the UB book scanner, only the colour setting (colour; b/w; grayscale) can be changed. The resolution cannot be changed (always 300dpi). The setting "Colour" and 300dpi is ideal for the UB book scanner as it has the lowest error rate, i. e. the fewest wrongly recognized characters during OCR conversion.
- With the NDL copier, both colour settings and resolution can be adjusted. Both parameters have been changed. The resolutions of 200,300,400 and 600 dpi were scanned with full colour (256 colours), b/w, grayscale and automatic colour control. The lowest error rate and thus the best result was achieved with a full colour setting of 400 dpi.

➔ The first choice is the NDL copier with Full Colour, 400dpi, because Abbyy makes 2 mistakes less on a double page. If you need to scan with the UB scanner, use the "Colour" setting (and 300dpi, which is not modifiable anyway).


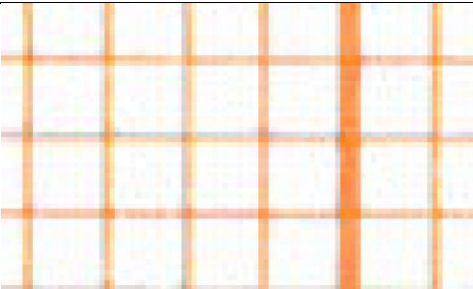


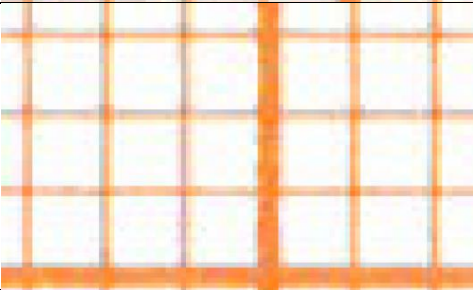

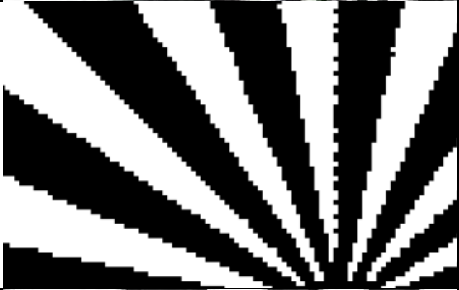
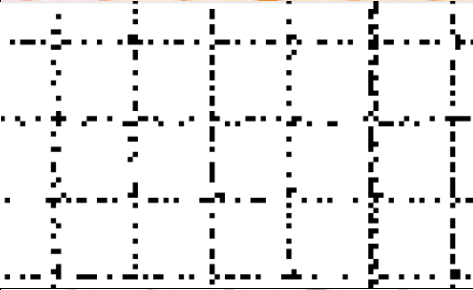













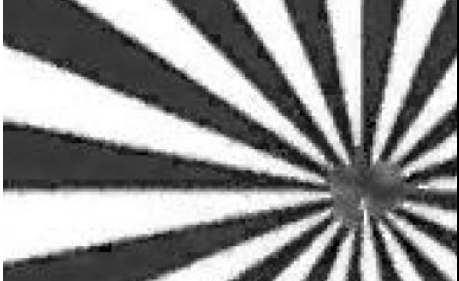
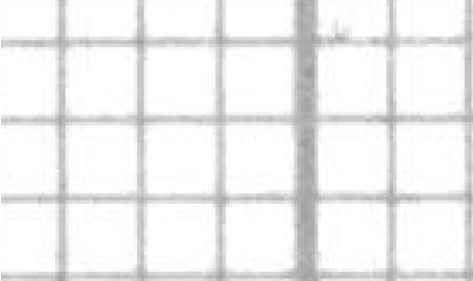

The table on the last page lists Abbyyy's conversion errors. You can clearly see that you can reduce the error rate to almost zero if you train Abbyy on the mistakes as a further step. I. e. the letter combination "st", which was usually recognized as "si", is trained as a character set, which means "st". See the 02 instructions for Abbyyy Fine Reader OCR of fracture texts. doc

UB book scanner	dpi	"Sun image"	Graph paper	Double page in fracture	Number of errors in the fracture text
colour	300				7
b/w	300				17
Grey levels	300				9



NDL-scanner	dpi	"Sun image"	Graph paper	Double page in fracture	Number of errors in the fracture text
full colour (256 colours)	200				7
Auto colour	200				7
b/w	200				22
Grey levels	200				8
full colour (256 colours)	300				8
Auto colour	300				8
b/w	300				9
Grey levels	300				8



full colour (256 colours)	400				5
Auto colour	400				6
b/w	400				10
Grey levels	400				6
(256 colours)	600				8
Auto colour	600				7
b/w	600				8
Grey levels	600				7



Error when converting from a double page to 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 book scanner																								
Farbe	300			1							1	1	1	2	1									
b/w	300	1		6						3						1	1	1	1	1				
Grey levels	300	1		5	1	1			1															
NDL scanner																								
Full colour	200			5			1			1														
Auto colour	200			5				1		1														
b/w	200	2	10	4		1		1		2			1											
Grey levels	200			5			1			1														
Full colour	300			4			1	2		1														
Auto colour	300			6																	2			
b/w	300			5				2		2														
Grey levels	300			5					1													2		
Full colour	400			5																				
Auto colour	400			5									1											
b/w	400			4			1	1		2													1	1
Grey levels	400			4			1	1																
Full colour	600			4			1	1		2														
Auto colour	600			4				1		2														
b/w	600			5			1			2														
Grey levels	600			4			1			2														

*training was only st, and tz

**Numbers were cut off

***every file has been retrained

****Time expenditure per double page:
train between 15 and 30 minutes
check between 5 and 10 minutes

¹ „→“stands for "instead of"