

KonDoc – endlich mit OCR

Der Campuslieferdienst kann jetzt lesen

Jana Wettstein

KonDoc – was ist das überhaupt?

KonDoc ist unser Campuslieferdienst – diesen Dienst bieten wir seit dem 24.10.2008 an. Mit KonDoc liefern wir Angehörigen der Universität Konstanz Artikel aus Zeitschriften und Büchern aus dem eigenen Bestand per E-Mail aus. Zehntausende von Aufträgen wurden über die Jahre mit KonDoc erledigt und ausgeliefert. Seit Ende 2025 auch wieder an Studierende! Dies war bisher nur kurzzeitig während der Corona-Pandemie möglich.

Ein paar Zahlen aus den letzten Jahren: 2013 hatten wir 25.100 Aufträge, 2016 hatten wir 19.429 Aufträge und 2024 waren es leider nur noch etwas über 7.000 Aufträge. Die Rückgänge spiegeln die veränderten Möglichkeiten, wie auf Literatur und digitale Bestände zugegriffen werden kann.

Um den Dienst wieder attraktiver zu gestalten und auch um mit der Zeit zu gehen, haben wir uns im Jahr 2024 dazu entschieden, den Dienst zu erweitern, indem wir unsere Scans mit OCR versehen.

OCR – was ist das überhaupt?

OCR steht für Optical Character Recognition – auf Deutsch optische Zeichenerkennung. Das heißt, die Buchstaben werden erkannt und ein Programm erzeugt daraus einen richtigen Text und legt diesen quasi hinter das „Bild“ des Textes. Hiermit ist der Text kopierbar und durchsuchbar. Dies verbessert die Datenqualität und die Automatisierung, unterstützt digitale Workflows und in unserem universitären Kontext vor allem auch das wissenschaftliche Arbeiten.

Jetzt denkt man sich: „Das ist ja einfach!“. So ging es zumindest mir: „Wir scannen Text ein, dieser kann doch einfach von dem Programm

gelesen werden und die OCR sofort umgesetzt werden.“

Ganz so einfach war es dann doch nicht. Bei der Konfiguration des Dienstes merkten wir recht schnell, dass dahinter ein sehr komplexes System steckt und man einige Voraussetzungen für eine gute OCR erfüllen muss. Zum einen brauchen wir eine gute Bildqualität – hohe Auflösung, scharfe Scans und gute Kontraste. Standardisierte Schriften sind ebenfalls von Vorteil. Des Weiteren sollten die Scans möglichst gerade ausgerichtet sein und der Text darf nicht abgeschnitten sein. Natürlich sind ordentliche und gute Scans auch unser eigener Anspruch, aber je nach Beschaffenheit, Alter und Bindung der Bücher ist das nicht immer ganz einfach.

Um ein möglichst gutes Ergebnis der OCR zu erzielen, muss sich tatsächlich fast jede einzelne Seite, welche gescannt wurde, nochmal genau angesehen werden. Für diese Sichtung werden die Scans zunächst von unserem Dokumentliefersystem myBib eDoc in ein anderes Programm heruntergeladen. Dort werden die Scans manuell gesichtet und nachbearbeitet.

Auf was achten wir?

Gibt es an den Rändern schwarze Streifen? In welcher Sprache ist der Text verfasst? Hat der Text mehrere Spalten? Ist alles gut lesbar?

Die Systeme, die hinter einer OCR stecken, brauchen möglichst genaue Informationen. Sehr wichtig ist z. B., dass die Software erkennt, in welcher Sprache der Text verfasst ist. Bei uns sind die Sprachen Deutsch, Englisch und Französisch als Standardangebot voreingestellt, da Bestellungen über KonDoc in der Regel Inhalte in deutscher, englischer und französischer Sprache umfassen. Zur Auswahl

haben wir aber einen bunten Strauß aller Sprachen dieser Welt. Würden wir aber dem Programm alle gleichzeitig anbieten, käme dieses durcheinander. Bemerken wir also, dass wir Bestellungen für Scans in anderen Sprachen erhalten, wählen wir die entsprechende Sprache zusätzlich aus. Des Weiteren müssen wir beachten, ob sich der Text über mehrere Spalten gliedert. Ist dies der Fall, muss dies entsprechend gekennzeichnet werden, sonst wird die Lücke zwischen den Spalten überlesen und man erhält eine zusammenhangslose OCR.

Und nun?

Wenn alles passt und alle Einstellungen an den Scans vorgenommen wurden, werden diese nun mit einer OCR versehen und wieder in unser Dokumentliefersystem myBib eDoc zurückgeschickt und dort vollautomatisiert weiter an die Bestellenden ausgeliefert.

Das klingt zunächst nach einer Menge Aufwand und Handarbeit. Zugegeben – der manuelle Anteil ist größer, als wir anfangs dachten. Insgesamt ist der Prozess aber gut machbar:

Viele Einstellungen sind in unserer Software bereits global hinterlegt, nur manche müssen noch manuell vorgenommen werden.

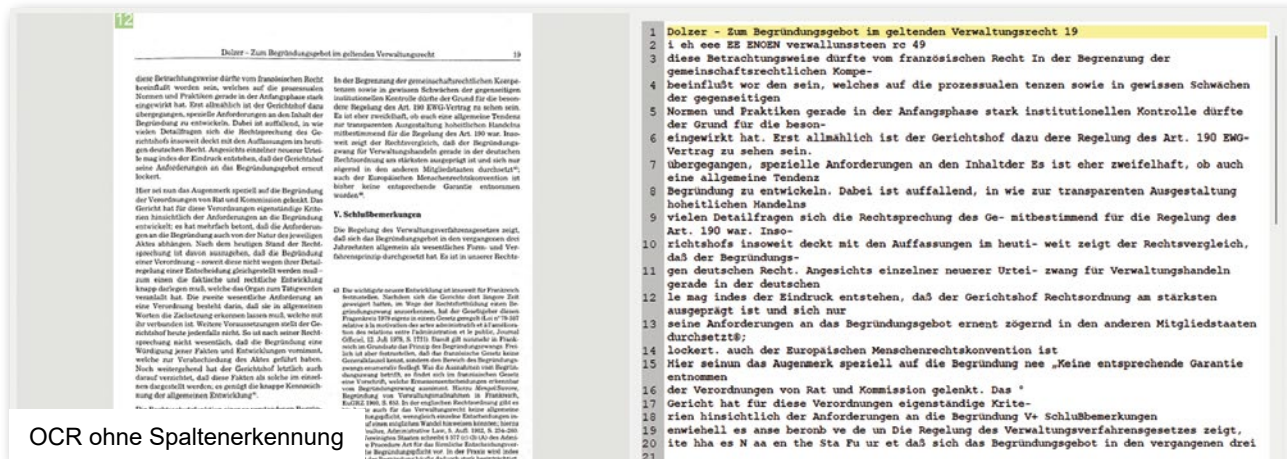
Sind alle glücklich damit?

Wir wissen es nicht – seit Einführung des Dienstes haben wir leider keinerlei Resonanz erhalten, weder positive noch negative. Wir werten das aber als gutes Zeichen! Im schwäbischen sagt man: Ned gschompfa isch globt g'nug!

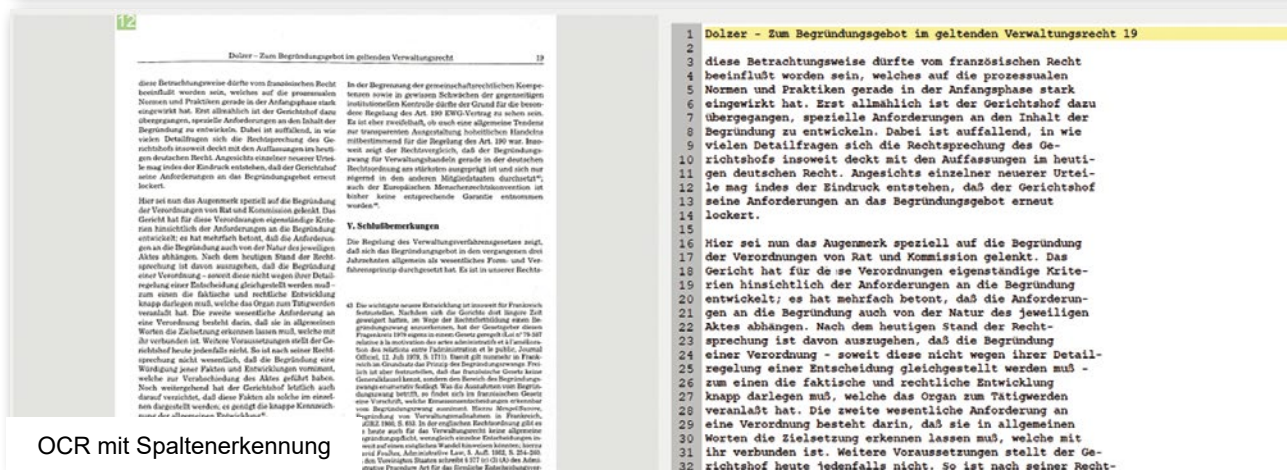
Und zum Schluss

Wir laden alle herzlich ein unser Angebot gerne zu nutzen! KonDoc bietet eine einfache und bequeme Möglichkeit, benötigte Materialien direkt zu erhalten – schnell, zuverlässig und unkompliziert.

Wie zu Beginn bereits erwähnt, freuen wir uns sehr, dass KonDoc seit Ende 2025 auch wieder Studierenden zur Verfügung steht. Damit können nun alle Hochschulangehörigen von unserem Service profitieren!



OCR ohne Spaltenerkennung



OCR mit Spaltenerkennung