LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

OCR
- Evaluierung der Genauigkeit (QM) sowie Tools zur Unterstützung

Manuela Schink

Online-Konferenz **„OCR – Prozesse und Entwicklungen“**
1. März 2021

Leibniz Gemeinschaft

Agenda mit Hintergrundgrafik

1. Ausgangslage zur OCR-Qualität
2. Wie und was machen andere Bibliotheken?
3. Setup für erste Tests
4. Ergebnisse
5. Ausblick
6. Diskussion

Ausgangslage zur OCR-Qualität

- Erste händische Überprüfung bei Start der Retrodigitalisierung 2014 und 2015
- Workshop 2020 "Effizienz und Qualitätssicherung in Digitalisierungsworkflows“

- **Aktuelle Überlegungen**:
  - Regelmäßige Qualitätsmessung (QM) notwendig?
  - Nachnutzung der OCR durch Metadaten-Erschließung möglich?
  - Gedankenspiel eigene OCR-Farm aufzubauen, bspw. mit Unterstützung von OCR-D

→ Kenntnisse über **OCR-Qualität** wichtig

→ **Komfortable Messmöglichkeiten** sollten gefunden werden

Wie und was machen andere Bibliotheken?

1. Workshop Retrodigitalisierung (2020)
2. Community

… nicht viel

… nicht regelmäßig

Setup für erste Tests

- 8 Digitalisate, antiqua
- 2 OCR-Versionen:
  - Produktiv-System = DL-OCR (Tesseract, V. 3), mit Pre-Processing
  - Test-System = Tesseract (V.4), roh

Händische Überprüfung:

a) 15 Wörter pro Digitalisat (120-150 Zeichen pro Buch)
   - gleiche Wörter aus beiden OCR

b) Auf einer Seite im Digitalisat alle Fehler markieren

Bernoulli: Überprüfung durch DL-eigenes Tool (500 Zeichen pro Buch)

Ergebnisse händische Überprüfung - Handhabung

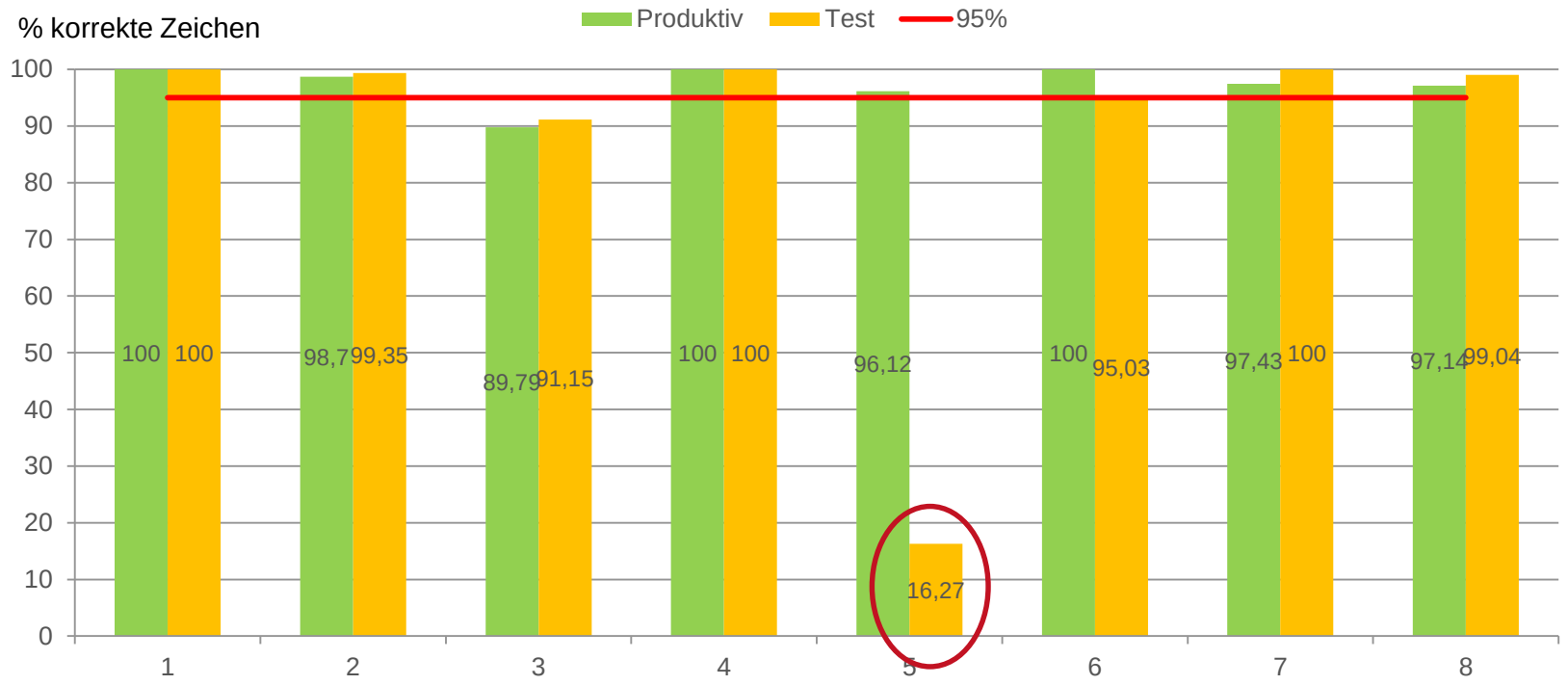
- Sehr hoher Aufwand bringt geringere Stichprobengröße
- hohe Fehleranfälligkeit, Auswahl wenig zufällig
- Fehlende Worte beeinflussen das Ergebnis stärker
- Keine Vorbereitung notwendig, dafür viel Nachbereitung
- Ergebnisse sind dokumentiert, nachvollziehbar

PROD							
Goobi-Vorgangstitel	**Bildnummer**	**Seite**	**Wort**	**Anzahl Zeichen/Wort**	**Anzahl korrekte Zeichen**	**Anzahl unrichtige Zeichen**	**Bemerkungen**
baukfudip_1713913984_12	11	[VI]	mustergültige	13	13		auseinander geschrieben
	14	[1]	Berufes	7	7		
	16	3	II. Planungsamt	14	0	14	nicht zu finden
	40	27	Bärgewicht	10	10		
	59	46	Gründung	8	7	1	Z statt G
	79	66	Müssen	6	6		
	111	98	Ausfüllung	10	10		
	177	164	stockweise	10	10		
	204	191	Stürmen	7	7	1	Stüurmen
	233	220	Hängewerk	9	9		
	279	266	Verkürzung	10	10		
	301	288	Kesselhäuser	12	12		
	329	316	Steindollen	11	11		
	377	364	Frostwetter	11	11		
	382	[369]	Außenputz	9	9		
gesamt				**147**	**132**	**16**	

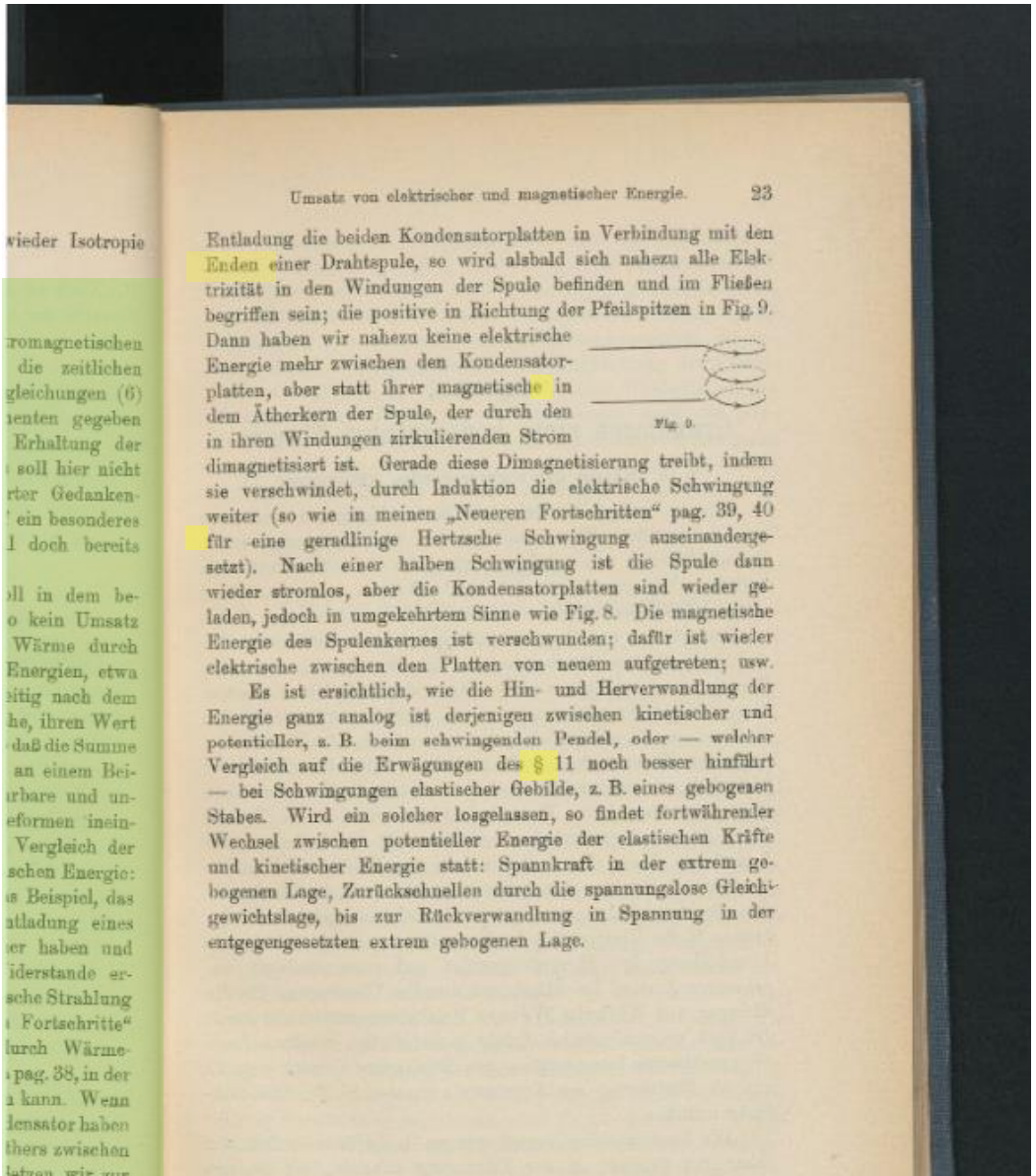
Ergebnisse händische Überprüfung - Auswertung

- Geprüfte Zeichen je System: 1111
- Erkennungsquote **Produktiv: 96,58%**
- Erkennungsquote **Test: 86,31%**

→ entspricht den Erfahrungen und Erwartungen



Händische Überprüfung – alle Fehler einer Seite



Umsatz von elektrischer und magnetischer Energie. 23

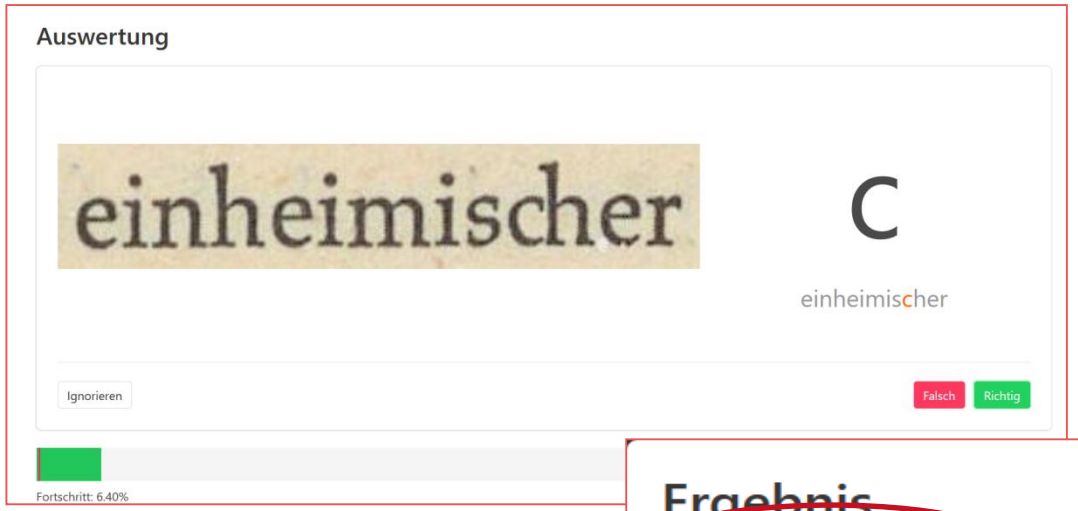
Entladung die beiden Kondensatorplatten in Verbindung mit den Enden einer Drahtspule, so wird alsbald sich nahezu alle Elektrizität in den Windungen der Spule befinden und im Fließen begriffen sein; die positive in Richtung der Pfeilspitzen in Fig. 9. Dann haben wir nahezu keine elektrische Energie mehr zwischen den Kondensatorplatten, aber statt ihrer magnetische in dem Ätherkern der Spule, der durch den in ihren Windungen zirkulierenden Strom dimagnetisiert ist. Gerade diese Dimagnetisierung treibt, indem sie verschwindet, durch Induktion die elektrische Schwingung weiter (so wie in meinen „Neueren Fortschritten" pag. 39, 40 für eine geradlinige Hertzsche Schwingung auseinandergesetzt). Nach einer halben Schwingung ist die Spule dann wieder stromlos, aber die Kondensatorplatten sind wieder geladen, jedoch in umgekehrtem Sinne wie Fig. 8. Die magnetische Energie des Spulenkernes ist verschwunden; dafür ist wieder elektrische zwischen den Platten von neuem aufgetreten; usw.

Fig. 9.

Es ist ersichtlich, wie die Hin- und Herverwandlung der Energie ganz analog ist derjenigen zwischen kinetischer und potentieller, z. B. beim schwingenden Pendel, oder — welcher Vergleich auf die Erwägungen des § 11 noch besser hinführt — bei Schwingungen elastischer Gebilde, z. B. eines gebogenen Stabes. Wird ein solcher losgelassen, so findet fortwährender Wechsel zwischen potentieller Energie der elastischen Kräfte und kinetischer Energie statt: Spannkraft in der extrem gebogenen Lage, Zurückschnellen durch die spannungslose Gleichgewichtslage, bis zur Rückverwandlung in Spannung in der entgegengesetzten extrem gebogenen Lage.

Umsatz von elektrischer und magnetischer Energie. 23
vieder Isotropie Entladung die beiden Kondensatorplatten in Verbindung mit den
Erden einer Drahtspule, so wird alsbald sich nahezu alle Elek-
trizität in den Windungen der Spule befinden und im Fließen
begriffen sein; die positive in Richtung der Pfeilspitzen in Fig. 9.
romagnetischen Dann haben wir nahezu keine elektrische
die zeitlichen Energie mehr zwischen den Kondensator- a
zleichungen. (6) platten, aber statt ihrer magnetische' in a
1enten gegeben dem Ätherkern der Spule, der durch den
Erhaltung der in ihren Windungen zirkulierenden Strom ;
; soll hier nicht dimagnetisiert ist. Gerade diese Dimagnetisierung treibt, indem
rter Gedanken- sie verschwindet, durch Induktion die elektrische Schwingung
' ein besonderes weiter (so wie in meinen „Neueren Fortschritten" pag. 39, 40
1 doch bereits für ,eine geradlinige Hertzsche Schwingung auseinanderge-
setzt). Nach einer halben Schwingung ist die Spule dann
ll in dem be- wieder stromlos, aber die Kondensatorplatten sind wieder ge-
o kein Umsatz laden, jedoch in umgekehrtem Sinne wie Fig. 8. Die magnetische
Wärme durch Energie des Spulenkernes ist verschwunden; dafür ist wieder
Energien, etwa elektrische zwischen den Platten von neuem aufgetreten; usw.
öitig nach dem Es ist ersichtlich, wie die Hin- und Herverwandlung der
he, ihren Wert Energie ganz analog ist derjenigen zwischen kinetischer und
daß die Summe potentieller, z. B. beim schwingenden Pendel, oder — welcher
an einem Bei- Vergleich auf die Erwägungen des $ 11 noch besser hinführt
rbare und un- — bei Schwingungen elastischer Gebilde, z. B. eines gebogenen
eformen "inein- Stabes. Wird ein solcher losgelassen, so findet fortwährender
Vergleich der Wechsel zwischen potentieller Energie der elastischen Kräfte
schen Energie: und kinetischer Energie statt: Spannkraft in der extrem. ge-
,s Beispiel, das bogenen Lage, Zurückschnellen durch die spannungslose Gleich"
atladung eines gewichtslage, bis zur Rückverwandlung in Spannung in der
er haben und entgegengesetzten extrem gebogenen Lage.
iderstande er-
sche Strahlung
Fortschritte"

Ergebnisse Bernoulli - Handhabung



https://ocrquality.goobi.io

- DL-Tool ist komfortabel zu nutzen, schnell
- abhängig von Kulanz des Mitarbeiters
- Zip-files müssen auf bestimmte Art strukturiert sein
- Tesseract-Version muss implementiert sein

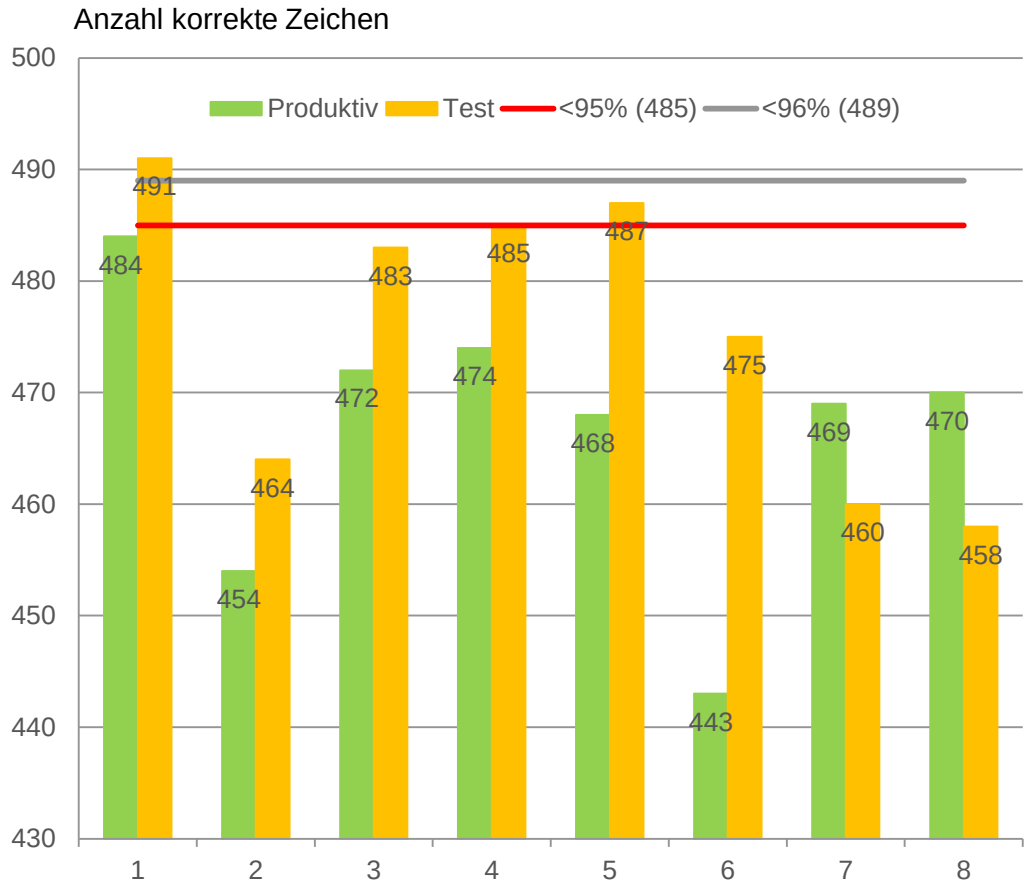
Ergebnis

484 richtig erkannt

Sie können das Ergebnis in der untenstehenden Tabelle einordnen.

Behauptete Erkennungsquote	Mindestzahl der korrekt erkannten Zeichen (Stichprobengröße=500)	
95 %	485	X
96 %	489	
97 %	493	
98 %	496	
99 %	499	

Ergebnisse Bernoulli - Auswertung



- Geprüfte Zeichen je System: 4000
- Erkennungsquote insgesamt
  - **Produktiv: 93,35%**
  - **Test: 95,08%**

→ Überraschung!

1. OCR Produktivsystem erfüllt nicht die Anforderungen für wissenschaftliche Arbeiten (95,5% Treffergenauigkeit)

2. Produktivsystem hat eine schlechtere Erkennungsquote als das Testsystem

3. Händische Überprüfung bringt ein deutlich anderes Ergebnis als Bernoulli

Mögliche Gründe für die Ergebnisse

Zu 2. Testsystem arbeitet mit neuerer Version von Tesseract

Zu 3. Händische Überprüfung ergibt deutlich bessere Ergebnisse

- Stichprobe bei der händischen Überprüfung deutlich kleiner (1111 Zeichen vs. 4000 Zeichen) → Größere Stichprobe findet mehr Fehler
- Zufällige Auswahl bei der händischen Überprüfung deutlich geringer → Mitarbeiterin hat sich selbst das Wort ausgesucht (“Schafft die OCR auch so ein schwieriges Wort? Wird die Kopfzeile erkannt?”)
- Art der Stichprobe “Wort pro Seite” verfälscht das Ergebnis, weil nur an wenigen Stellen überprüft wird

- Zu 1. ???

Ausblick

OCR-QM bleibt bei uns Thema !

- Weiterere Tests mit größerer Menge an Werken
- ggf. homogenere Testmenge (Sprache, Art)
- Tests von anderen OCR-QM-Tools
- Überprüfung der Scans auf Faktoren, die die OCR-Qualität mindern

→ Ohne OCR-QM-Tools keine regelmäßige/standardmäßige OCR-QM

- Händische Überprüfung zu aufwändig und Fehleranfällig (erster Eindruck)
- Festlegung auf ein Tool zur Überprüfung der OCR
- Eigene OCR mit Preprocessing

Vielen Dank für Ihre Aufmerksamkeit!

Fragen zur Diskussion

Welche Erfahrungen bei der OCR-QM gibt es bei Ihnen?

Ist eine standardisierte OCR-QM (im Workflow) sinnvoll trotz des hohen Aufwands?

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

MEHR INFORMATIONEN

www.tib.eu

Kontaktdaten
Manuela Schink
T 0511 762-9346, manuela.schink@tib.eu

Leibniz Gemeinschaft