

Umsetzungsbeispiel OCR

Beschreibung des Artefakts

Optical Character Recognition (OCR) ist eine Softwaretechnologie, die Texte in Bildern oder gescannten Dokumenten automatisch erkennt. Die Software analysiert digitale Bilddaten und identifiziert darin enthaltene Buchstaben, Zahlen und Zeichen, um sie in maschinenlesbaren und bearbeitbaren Texten umzuwandeln. Dadurch können beispielsweise gescannte Dokumente, PDFs oder Fotos von historischen Quellen digital durchsucht, kopiert oder weiterverarbeitet werden. Die OCR-Software wird in der Dokumentendigitalisierung, Archivierung und Datenverarbeitung eingesetzt und ermöglicht eine effizientere Nutzung von zuvor nicht digital zugänglichen Informationen bzw. Quellen.

Beschreibung des Szenarios

Den Schülerinnen und Schülern werden eingescannte Seite aus einer historischen Quelle zum Schafrichter in Dresden zur Verfügung gestellt. Die historische Quelle liegt zunächst nur als Bild vor, damit die Lernenden mithilfe einer OCR-Software in bearbeitbaren digitalen Text umwandeln.

Im nächsten Schritt überprüfen die Schülerinnen und Schüler das OCR-Ergebnis und korrigieren mögliche Erkennungsfehler. Dadurch setzen sie sich intensiver mit der historischen Quelle auseinander und lernen zugleich, typische Probleme der digitalen Texterkennung zu erkennen.

Durch dieses Vorgehen verbinden die Lernenden fachliche Kompetenzen der historischen Quellenanalyse mit digitalisierungsbezogene Kompetenzen im Umgang mit modernen Technologien zur Textverarbeitung und Digitalisierung historischer Materialien.

Basisangaben

Fach	Geschichte	Klassenstufe	Klasse 7
Schulform	Gymnasium	Lernbereich	LB 2
Zeitbedarf	90 min	Kompetenzrahmen	

Artefaktsteckbrief OCR

Wofür wird es genutzt?

Ein häufiges Einsatzgebiet von OCR-Software ist die Digitalisierung von historischen Quellen in Bibliotheken oder Archiven. Papierdokumente, Bücher oder Zeitungen können eingescannt und mithilfe von OCR in durchsuchbare Dateien umgewandelt werden. Dadurch lassen sich große Mengen an Informationen effizient speichern und mithilfe von Datenbanken leichter wiederfinden.

Auch in der Verwaltung kann OCR genutzt werden, beispielsweise um Rechnungen, Formulare oder Verträge automatisch zu erfassen und wichtige Daten digital weiterzuverarbeiten.

Darüber hinaus wird OCR in der Forschung verwendet, etwa um historische Quellen oder alte Bücher zu digitalisieren und für Analysen zugänglich zu machen.

Wie funktioniert es?

OCR funktioniert auf Grundlage der Mustererkennung (Pattern Recognition). Dabei analysiert eine Software Bilder oder Scans von Dokumenten und vergleicht die darin enthaltenen Formen von Buchstaben, Zahlen und Satzzeichen mit bekannten Schriftmustern in einer internen Datenbank. Anschließend setzt das Programm die erkannten Zeichen zu Wörtern und Sätzen zusammen und wandelt sie in digitalen, bearbeitbaren Text um.

Der Prozess der Texterkennung erfolgt in mehreren Schritten. Zunächst wird das Layout untersucht und die Software erkennt Textbereiche, Tabellen oder Grafiken und bereitet das Bild für den nächsten Schritt vor. In der Segmentierung werden anschließend Textteile von Bildern oder anderen grafischen Elementen getrennt. Danach folgt die Zeichenerkennung, bei der einzelne Buchstaben und Zahlen mithilfe von Mustererkennung identifiziert und mit gespeicherten Schriftmustern abgeglichen werden. Bei neuerer OCR-Software wird in der Nachbearbeitung mögliche Fehler durch intelligente Verfahren korrigiert, das sogenannte Intelligent Character Recognition (ICR) oder künstliche Intelligenz, die den Kontext von Wörtern und Sätzen analysiert. Verwechslungen, wie z. B. „O“ und „0“ erkennt die künstliche Intelligenz und korrigiert diese. Abschließend wird der erkannte Text in bearbeitbare Formate wie Word-, Excel- oder durchsuchbare PDF-Dateien ausgegeben.

Durch den Einsatz von künstlicher Intelligenz kann moderne OCR-Software außerdem aus großen Mengen an Textdaten lernen und ihre Erkennungsgenauigkeit kontinuierlich verbessern. Dadurch ist es möglich auch handschriftliche Dokumente zuverlässig zu digitalisieren.

Umsetzung im Unterricht

Technisch-/ infrastrukturelle Voraussetzungen

Die Schülerinnen und Schüler benötigen einen Zugang zur OCR-Software „Transkribus“, die über das Internet erreichbar ist. Daher werden digitale Endgeräte, beispielsweise Computer, Laptops oder Tablets benötigt. Nachteil bei der Software „Transkribus“ ist, dass man ohne Anmeldung nur einzelne Bilder transkribiert werden können. Durch eine kostenlose Anmeldung können ganze Projekte transkribiert werden, dennoch ist die kostenlose Nutzung eingeschränkt und kann längere Wartezeiten mit sich bringen.

Darüber hinaus müssen die zu bearbeitenden Materialien in digitaler Form vorliegen, als digitale Fotografien. Diese können in der digitalen Lernumgebung „LernSax“ bereitgestellt werden und Ergebnisse ausgetauscht werden.

Curriculare Analyse

Fachlehrplan

In dem Fach Geschichte werden in der gesamten Schulzeit mit verschiedene historische und aktuelle Quellenarten gearbeitet und umgegangen. Dafür ist im Lehrplan genau aufgeführt welche Quellenart wann und in welcher Epoche behandelt werden sollte. Dabei wird aber auf keiner digitalisierungsbezogene Kompetenzen hingewiesen.

Das Thema „Schafrichter“ kann in Klasse 7 LB 2 eingeordnet werden, da hier die Gesellschaft beleuchtet wird und ein regionales Ereignis dazu benötigt wird. Ebenso soll die Quellenart „Schriftliche Quellen“ betrachtet und erklärt werden.

digitalisierungsbezogene Lernziele

#	MIA	K#	Qualifikationsebene Kompetenzrahmen	Handlungsebene
D1	M	1.2.1	Die SuS analysieren, interpretieren und bewerten kritisch Informationen und Daten	indem sie die transkribierten Texte sinnvoll und grammatikalisch richtig korrigieren.
D2	A	3.1.1	Die SuS kennen und wenden mehrere technische Bearbeitungswerkzeuge für digitale Medien an	indem sie eigenständig die OCR-Software bedienen und ihren Partnern die Vorgehensweise erklären.
D3	I	3.2.2	Die SuS verarbeiten Informationen, Inhalte und vorhandene digitale Produkte weiter und integrieren diese in bestehendes Wissen	indem sie die bereitgestellten Bilder in „Transkribus“ hochladen und das Ergebnis präsentieren.
D4	A	5.2.3	Die SuS identifizieren passende Werkzeuge zur Lösung	indem sie die richtige OCR-Software in „Transkribus“ auswählen.
D5	M	6.2.2	Die SuS erkennen Chancen und Risiken des Mediengebrauchs in unterschiedlichen Lebensbereichen, können eigenen Mediengebrauch reflektieren und ggf. modifizieren	indem sie die Grenzen der OCR-Software kennen und die transkribierten Texte korrigieren.

Unterrichtsverlaufsplanung

Legende		Unterrichtsmedium	Unterrichtswerkzeug	Unterrichtsgegenstand		
Zeit	Phase	Inhalt	SF/M ¹	Medien/Material	LZ Nr.	
10'	Einstieg	Impulsbild: L: Beschreibt das Bild. Identifiziert Schwierigkeiten, die beim Lesen entstehen können. Stellt Vermutungen an, wie die Arbeitsweise von Historikern bei der Arbeit mit solchen Quellen aussehen kann. SuS: -andere Schriftart, schwierig zu entziffern, Handschrift	UG	Beispiel Bild		

¹SF – Sozialform (EA: Einzelarbeit, PA: Partnerarbeit, GA: Gruppenarbeit; FU: Frontalunterricht); M - Methode

Zeit	Phase	Inhalt	SF/M ¹	Medien/Material	LZ Nr.
5'	Einführung	Erklärung OCR-Software: OCR= Optical Character Recognition Erkennt Text in Bildern und wandelt diese in digitale Texte um Wird angewendet um alte Schriftarte, besonders Handschriften, schnell zu transkribieren Arbeitet anhand von „gelernten“ Mustern Kann aber auch Fehler machen: z.B. „0“ anstatt einem „O“ -> zurückzuführen auf Muster Am Ende muss der Text nochmal korrigiert werden. Erklärung „Transkribus“	FU	PPP	D2
45'	Erarbeitung	SuS arbeiten mit Partner zusammen an 1-3 Blätter der bereitgestellten Akten Bearbeitung in der Software „Transkribus“ Text soll transkribiert und im Anschluss korrigiert werden	PA	Bereitgestellte Quellen, Transkribus	D1, D2, D3, D4

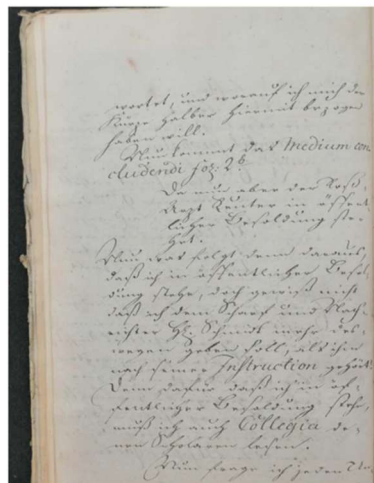
Zeit	Phase	Inhalt	SF/M ²	Medien/Material	LZ Nr.
20'	Ergebnissicherung	Partnergruppen präsentieren, in der richtigen Reihenfolge, ihren transkribierten Text.	UG	Transkribierte Texte	D1, D3
		Erkennen typischer Fehler der OCR			
		Beschreibt die Aufgaben, die ein Schafrichter in Dresden hatte.			
10'	Reflexion	Beschreibt die Grenzen und Chancen der OCR bei der Arbeit von Historikern.	UG		D5

²SF – Sozialform (EA: Einzelarbeit, PA: Partnerarbeit, GA: Gruppenarbeit; FU: Frontalunterricht; UG: Unterrichtsgespräch); M - Methode

Arbeitsmaterialien

OCR-SOFTWARE

Wie Computer alte Texte lesen können



WAS IST OCR?

- OCR = Optical Character Recognition
- Eine Software erkennt Buchstaben in einem Bild anhand von Mustern
- Digitaler Text wird ausgegeben
- Texte können durchsucht, kopiert und besser untersucht werden
- Oftmals Handschriften schwer zu entziffern

WARUM IST DAS FÜR GESCHICHTE WICHTIG?

- OCR hilft dabei, alte Texte schneller zu lesen
- Schrift muss nicht aufwendig gelernt werden
- Fehlerquote kann gesenkt werden



VORSICHT!

- Software macht manchmal Fehler
- Zum Beispiel kann ein 'O' als '0' erkannt werden
- Deshalb müssen Menschen die Texte immer kontrollieren



WAS IST TRANSKRIBUS?

- Transkribus ist eine spezielle OCR-Software
- Sie hilft Forschern und Schülern, alte Texte zu lesen
- Besonders gut erkennt sie historische Handschriften

WIE FUNKTIONIERT TRANSKRIBUS?

1. Eine Transkriptionsprogramm wird ausgewählt
2. Ein Foto oder Scan eines Dokuments wird hochgeladen
3. Die Software erkennt Buchstaben und Wörter
4. Der Text und Layout wird automatisch dargestellt
5. Überprüft den Text und korrigiert Fehler

