109. Bibliothekartag
buten un binnen – wagen un winnen
16.-18.06.2021 | Bremen

GEORG ECKERT INSTITUTE
for International Textbook Research

OCR ON DEMAND
DER ANSATZ EINES USER GENERATED CONTENT

DR. ANKE HERTLING, SEBASTIAN KLAES

Member of the Leibniz Association

AUFGABEN GEORG-ECKERT-INSTITUT (GEI)

- internationale und multidisziplinäre **Schulbuch- und Bildungs-medienforschung** mit kulturwissenschaftlich-historischem Schwerpunkt: **Produktion, Inhalt und Rezeption** von Schulbüchern und Bildungsmedien

- **Berater** und **Mediator** in Schulbuch-Angelegenheiten

- Förderung der Schulbuch- und Bildungsmedienforschung durch Unterhalt einer öffentlich zugänglichen **Forschungsbibliothek** als zentrale **Forschungsinfrastruktur**

FORSCHUNGSBIBLIOTHEK GEI

Schulbücher und Lehrpläne der Fächer **Geschichte, Geographie, Politik, Sozialkunde, Ethik/Religion/Werteerziehung und (Erst)-Lesebücher**

- über 186.000 Schulbücher und Lehrpläne aus mehr als 175 Ländern
- davon 24.000 historische Bände
- wissenschaftliche Literatur

Einbindung Community

- Projektpartner Fachinformationsdienst (FID) Erziehungswissenschaft und Bildungsforschung
- OCR-D

Einbindung Wissenschaft

- Digital Humanities Drittmittelprojekte
- Forschungsschwerpunkt GEI Digitalität / Digital Humanities

/GeorgEckertInstitut @GeorgEckert

gei.digital

Die digitale Schulbuch-Bibliothek

Startseite **Suchen** Stöbern Aktuelles Über das Projekt Partner Nutzungsbedingungen OAI-Schnittstelle **DE / EN**

Sortierung

- Personen
- Titel
- Importdatum
- Erscheinungsjahr

Sammlungen

- Lesebücher Kaiserreich ✖

Verlag

- Oldenbourg (21)
- Crüwell (14)
- Hirt (10)

Sie befinden sich hier: Startseite Lesebücher Kaiserreich

Suche

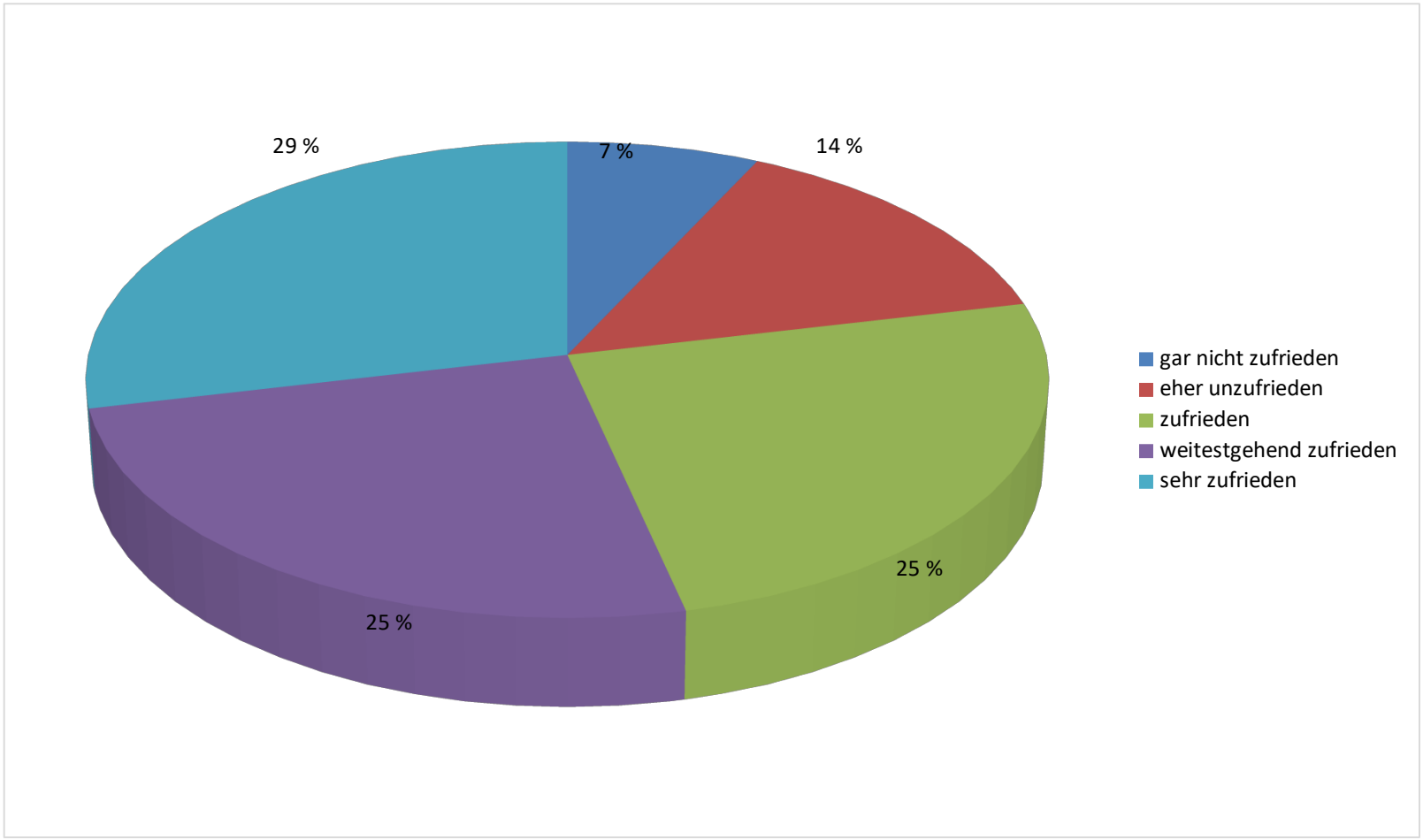
Treffer: 305

Standardsuche | Erweiterte Suche

in Werken der Sammlung: Lesebücher Kaiserreich

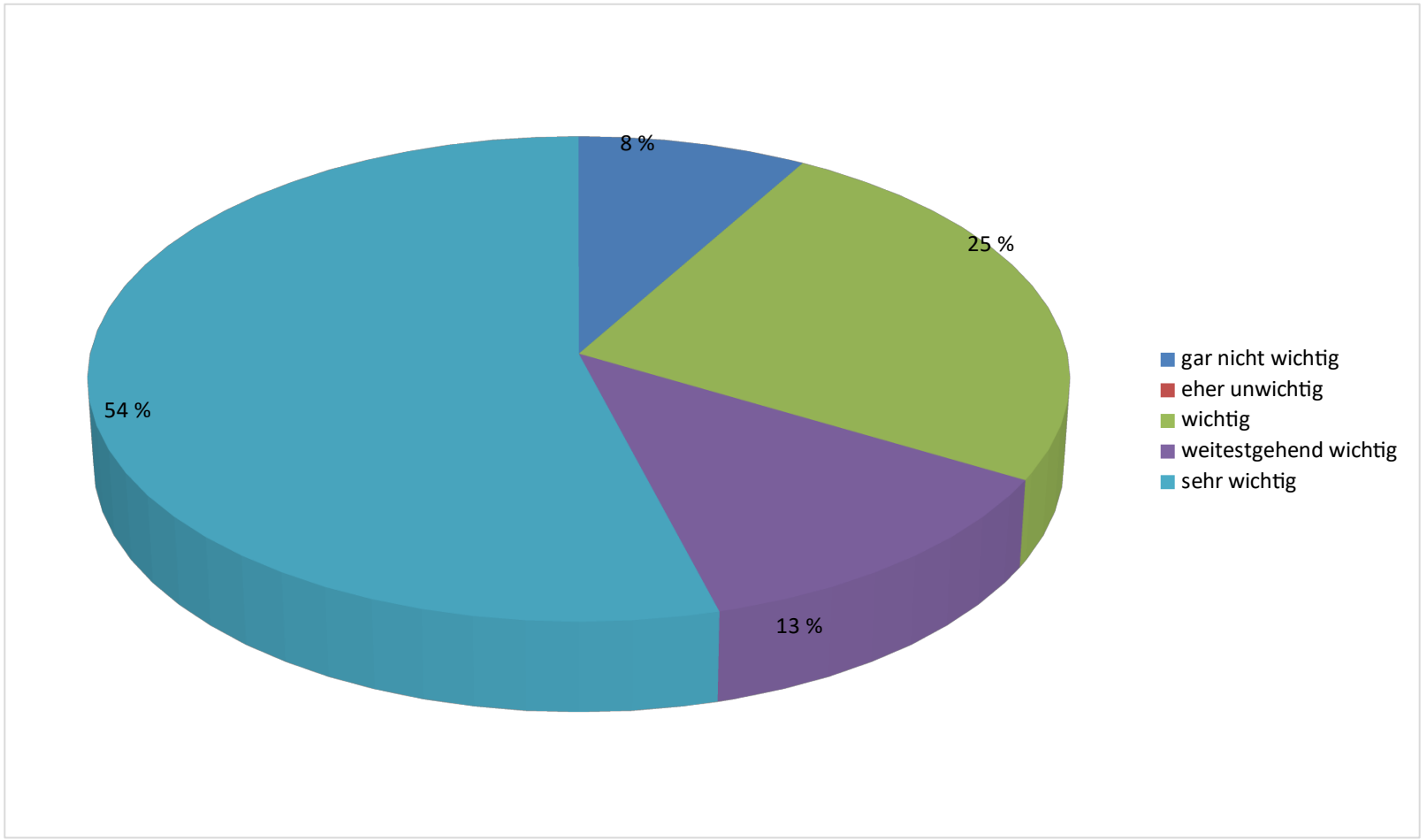
Allen Daten **Metadaten** **Volltexten**

USER SURVEY: ZUFRIEDENHEIT VOLLTEXTQUALITÄT



/GeorgEckertInstitut @GeorgEckert

USER SURVEY: EXPORTMÖGLICHKEIT VON VOLLTEXTEN (z.B. TextGrid etc.)



/GeorgEckertInstitut @GeorgEckert

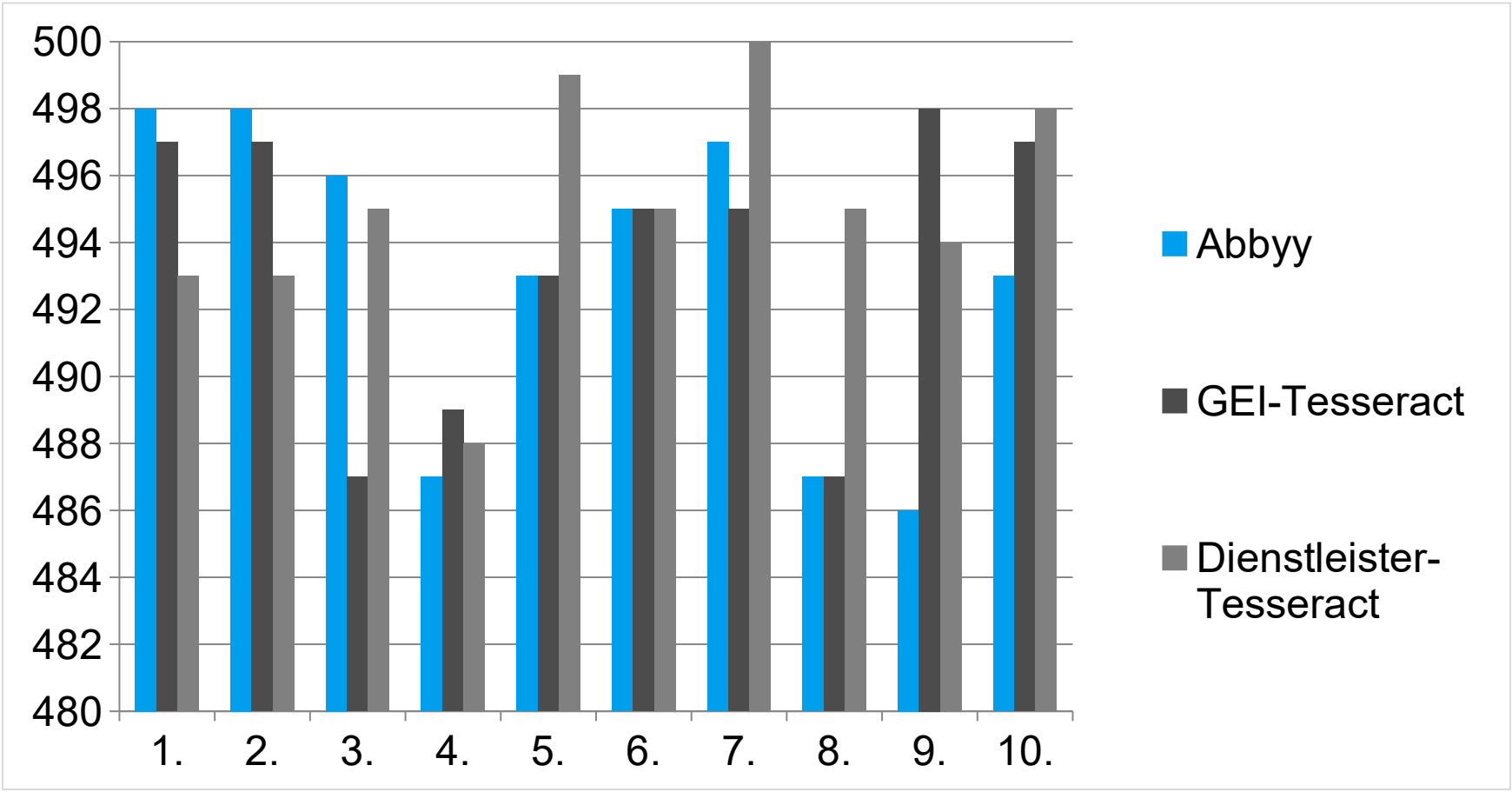
QUALITÄTSMESSUNG OCR

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

DFG-Praxisregeln: „Bernoulli-Experiment“

/GeorgEckertInstitut @GeorgEckert

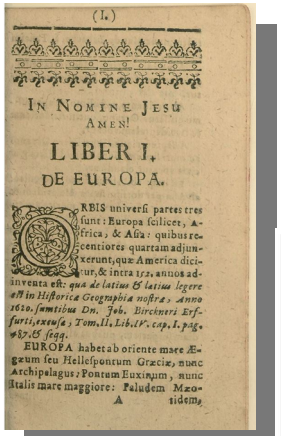
ERSTE ERGEBNISSE 2019



/GeorgEckertInstitut @GeorgEckert

USER NEEDS

- Basic: Volltextsuche
- Frequenzanalyse
- Kollokationsanalyse
- ...



IN NOMINE JESU
AMEN!

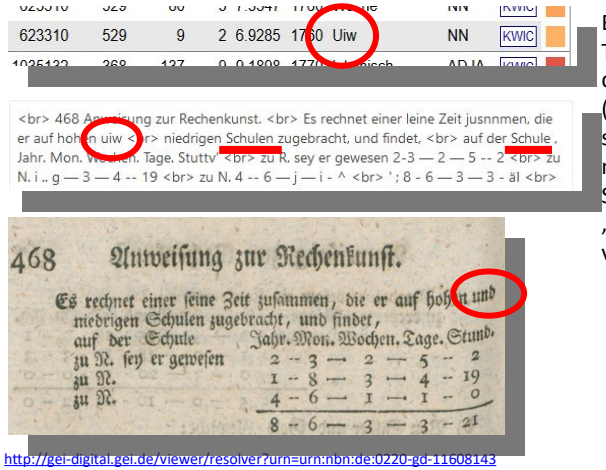
LIBER I.
DE EUROPA.

RBIS univerſi partes tres
ſunt: Europa ſcilicet, A-
frica, & Aſia: quibus re-
centiores quartam adjun-
xerunt, quæ America dici-
tur, & intra 1/3. annos ad-
inventa eſt: qua de latine & latine legere
eſt in Hiſtorica Geographia noſtra, Anno
1620. ſumtibus Dn. Joh. Birckneri Erf-
furti, exeuſa; Tom. II. Lib. IV. cap. I. pag.
987. & ſeqq.
EUROPA habet ab oriente maæ Æ-
gæum ſeu Helleſpontum Grzcix, nunc
Archipelagus; Pontum Euxinum, nunc
Italis mare maggiore: Paludem Mæo-
A tidem.

Beispiel 1:
Die früheste mit den digitalen Werkzeugen auffindbare Nennung von „Europa“ im Korpus stammt aus dem *Mercurius Cosmicus* von 1648. Die Nennungen von Afri[k]a, Asien und Ameri[k]a auf derselben Seite werden aufgrund mangelhafter OCR-Ergebnisse/lateinischer Schreibweise nicht gefunden/erkannt.

 (Ю <br> o O_O O O O O o oo o
 I ^ Nomine Jesu
 Am en!
 LIBER L
 DE EUROPA.
 RBIS univcrii partes tres <br> funt: Europa fcilicet, A-
 frica> &AG'a: quibus rc-
 centiores quartam adjun-
 xerunt,quae America dici¬
 tur,& intra i/j, anno* ad-
 inventa eft; qua ¿e latitu & Utua legere
 в Я in Hift or i c л Geograf bi л noftra , y.inno <br> iCio.fumtibu* Dn. Job. Birckneri Erf-
 furti,ex tu fa , Tomoli» Lib.IK, cap<, I. pag, <br> feqq.
 EUROPA habet ab oriente ma-re ЛЕ-
 gzum feu Hellefpontum Grzcix; mmc
 Archipelagus : Pontum Euxinam , nunc
 Stalis шare maggiore: Paludem M*o«
 A tidem#

http://gei-digital.gei.de/viewer/resolver?urn=urn:nbn:de:0220-gd-12794530



623310	529	9	2	6.9285	1760	Uiw	NN

 468 Anweiſung zur Rechenkunſt.
 Es rechnet einer leine Zeit jusnmen, die er auf hohen uiw <br> niedrigen Schulen zugebracht, und findet, <br> auf der Schule , Jahr. Mon. Wochen. Tage. Stuttv <br> zu R. sey er gewesen 2-3 — 2 — 5 -- 2 <br> zu N. i .. g — 3 — 4 -- 19 <br> zu N. 4 -- 6 — j — i - ^
 ' ; 8 - 6 — 3 — 3 - äl

468 Anweiſung zur Rechenkunſt.

Es rechnet einer ſeine Zeit zuſammen, die er auf hohen und niedrigen Schulen zugebracht, und findet,

auf der Schule	Jahr.	Mon.	Wochen.	Tage.	Stund.
zu N. ſey er geweſen	2	3	2	5	2
zu N.	1	8	3	4	19
zu N.	4	6	1	1	0
	8	6	3	3	21

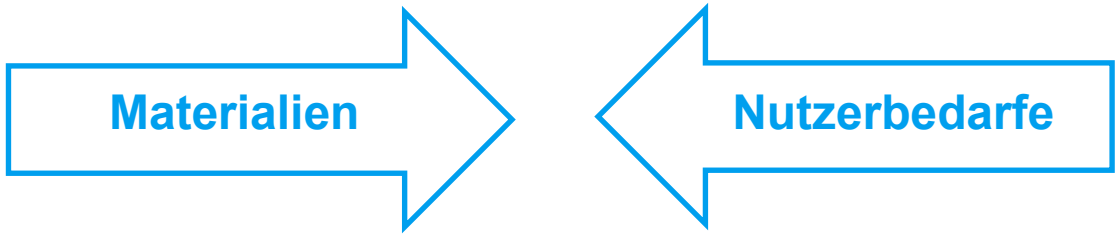
http://gei-digital.gei.de/viewer/resolver?urn=urn:nbn:de:0220-gd-11608143

Beispiel 2:
Tools interpretieren den OCR-Fehler „uiw“ (statt „und“) als seltenes, aber häufig mit dem gesuchten Stichwort (hier: „Schule“) vorkommendes Wort.

/GeorgEckertInstitut @GeorgEckert

OCR ITERATIV UND PARTIZIPATIV

- Texterkennung als Daueraufgabe und iterativer Prozess
- DH-Forschung und -Bedarfe stärker in OCR-Entwicklungen einbeziehen
- granularer und agiler Umgang mit Texterkennungsprozessen



- GroundTruth für OCR-Training und Trainingsmodelle breiter zur Verfügung stellen

/GeorgEckertInstitut @GeorgEckert

OCR4ALL LIBRARIES: USER GENERATED CONTENT

Modulprojekt in der DFG-Förderlinie e-Research-Technologien
„Implementierung der OCR-D-Software zur Volltextdigitalisierung“
(3. Förderphase 2021-2023)

Ziel DFG: Volltexttransformation der im deutschen Sprachraum erschienen Drucke 1600 bis 1850

Projektpartner

- Zentrum für Philologie und Digitalität (ZPD), Universität Würzburg
- Lehrstuhl für Mensch-Computer Interaktion (HCI), Universität Würzburg
- GEI-Forschungsbibliothek

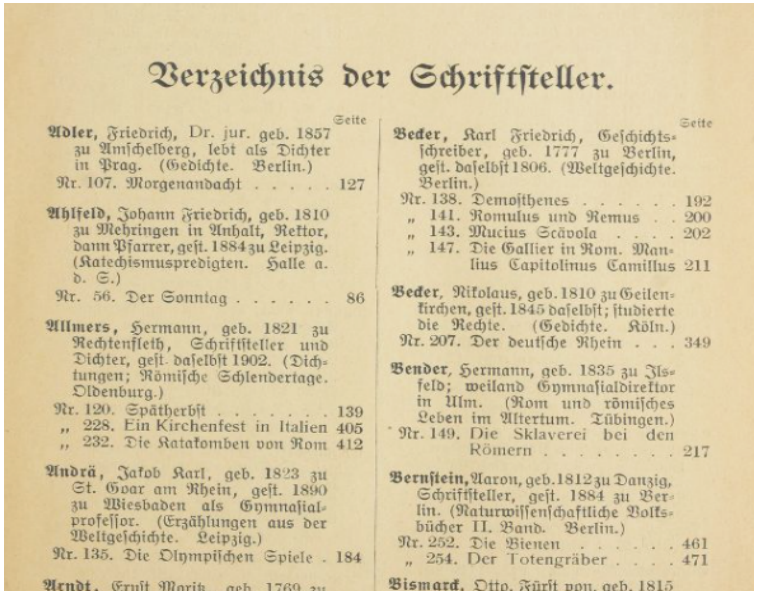
Projektziel: Implementierung **OCR-D-Komponenten** und Weiter-entwicklung der Open-Source-Software **OCR4all** für **niederschwellige, flexible und eigenständige** OCR-Massenverarbeitung

/GeorgEckertInstitut @GeorgEckert

OCR4ALL LIBRARIES: USER GENERATED CONTENT

Fokus Material: Organisation OCR-Optimierung entlang Materialspezifika

- Erzeugung von ersten spezifischen Trainingsdaten auf Basis von historischen Schulbüchern
- Training von OCR-Modellen bei historischen Schulbüchern



Verzeichnis der Schriftsteller.

Seite

Adler, Friedrich, Dr. jur. geb. 1857 zu Amschelberg, lebt als Dichter in Prag. (Gedichte. Berlin.)
Nr. 107. Morgenandacht 127

Ahlfeld, Johann Friedrich, geb. 1810 zu Mehringen in Anhalt, Rektor, dann Pfarrer, gest. 1884 zu Leipzig. (Katechismuspredigten. Halle a. d. S.)
Nr. 56. Der Sonntag 86

Allmers, Hermann, geb. 1821 zu Rechtenfleth, Schriftsteller und Dichter, gest. daselbst 1902. (Dichtungen; Römische Schlendertage. Oldenburg.)
Nr. 120. Spätherbst 139
„ 228. Ein Kirchenfest in Italien 405
„ 232. Die Katakomben von Rom 412

Andrä, Jakob Karl, geb. 1823 zu St. Goar am Rhein, gest. 1890 zu Wiesbaden als Gymnasialprofessor. (Erzählungen aus der Weltgeschichte. Leipzig.)
Nr. 135. Die Olympischen Spiele . 184

Arndt, Ernst Moritz, geb. 1769 zu

Seite

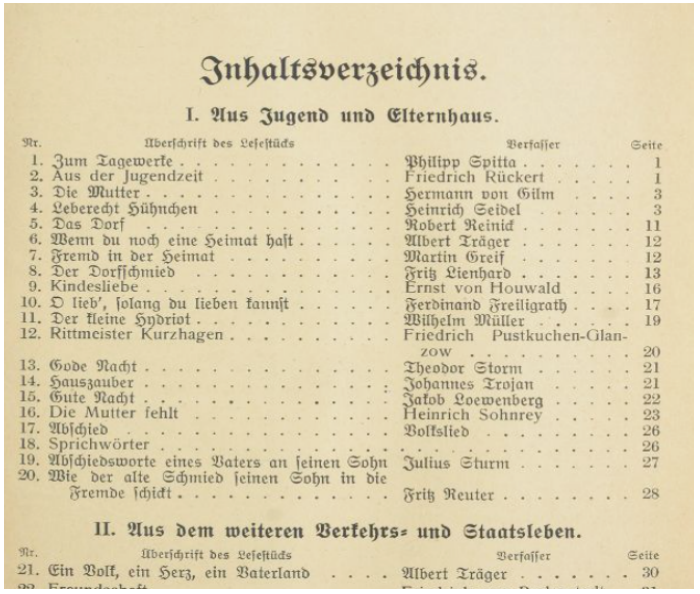
Becker, Karl Friedrich, Geschichtsschreiber, geb. 1777 zu Berlin, gest. daselbst 1806. (Weltgeschichte. Berlin.)
Nr. 138. Demosthenes 192
„ 141. Romulus und Remus . . 200
„ 143. Mucius Scävola 202
„ 147. Die Gallier in Rom. Manlius Capitolinus Camillus 211

Becker, Nikolaus, geb. 1810 zu Geilenkirchen, gest. 1845 daselbst; studierte die Rechte. (Gedichte. Köln.)
Nr. 207. Der deutsche Rhein . . . 349

Bender, Hermann, geb. 1835 zu Ilsfeld; weiland Gymnasialdirektor in Ulm. (Rom und römisches Leben im Altertum. Tübingen.)
Nr. 149. Die Sklaverei bei den Römern 217

Bernstein, Aaron, geb. 1812 zu Danzig, Schriftsteller, gest. 1884 zu Berlin. (Naturwissenschaftliche Volksbücher II. Band. Berlin.)
Nr. 252. Die Bienen 461
„ 254. Der Totengräber 471

Bismarck, Otto, Fürst von, geb. 1815



Inhaltsverzeichnis.

I. Aus Jugend und Elternhaus.

Nr.	Überschrift des Lesestücks	Verfasser	Seite
1.	Zum Tagewerke	Philipp Spitta	1
2.	Aus der Jugendzeit	Friedrich Rückert	1
3.	Die Mutter	Hermann von Gilm	3
4.	Leberecht Hühnchen	Heinrich Seidel	3
5.	Das Dorf	Robert Reinick	11
6.	Wenn du noch eine Heimat hast	Albert Träger	12
7.	Fremd in der Heimat	Martin Greif	12
8.	Der Dorfschmied	Fritz Lienhard	13
9.	Kindesliebe	Ernst von Houwald	16
10.	O lieb', solang du lieben kannst	Ferdinand Freiligrath	17
11.	Der kleine Hydriot	Wilhelm Müller	19
12.	Rittmeister Kurzhagen	Friedrich Pustkuchen-Glanzow	20
13.	Gode Nacht	Theodor Storm	21
14.	Hauszauber	Johannes Trojan	21
15.	Gute Nacht	Jakob Loewenberg	22
16.	Die Mutter fehlt	Heinrich Sohnrey	23
17.	Abschied	Volkslied	26
18.	Sprichwörter		26
19.	Abschiedsworte eines Vaters an seinen Sohn	Julius Sturm	27
20.	Wie der alte Schmied seinen Sohn in die Fremde schickt	Fritz Reuter	28

II. Aus dem weiteren Verkehrs- und Staatsleben.

Nr.	Überschrift des Lesestücks	Verfasser	Seite
21.	Ein Volk, ein Herz, ein Vaterland	Albert Träger	30

Lesebücher Deutsches Kaiserreich

/GeorgEckertInstitut @GeorgEckert

OCR4ALL LIBRARIES: USER GENERATED CONTENT

Fokus User (Wissenschaftler*innen, Bibliotheken): OCR4all Benutzeroberfläche GUI für nicht-technische Anwender*innen

- Entwicklung visuelle Erklärungskomponente zur Erstellung und Konfiguration optimaler OCR-Workflows
- Usability-Studien

Bild © ZPD http://www.ocr4all.org/de/home.php

/GeorgEckertInstitut @GeorgEckert

OCR4ALL LIBRARIES: USER GENERATED CONTENT

Fokus User (Bibliotheken): Konzept für Anbindung von OCR4all-libraries an Digitalisierungsworkflowsysteme

Interesse an weiteren Use Cases in Forschungs- und Spezialbibliotheken

Kontakt: gei-digital@gei.de

VIELEN DANK FÜR DIE AUFMERKSAMKEIT!