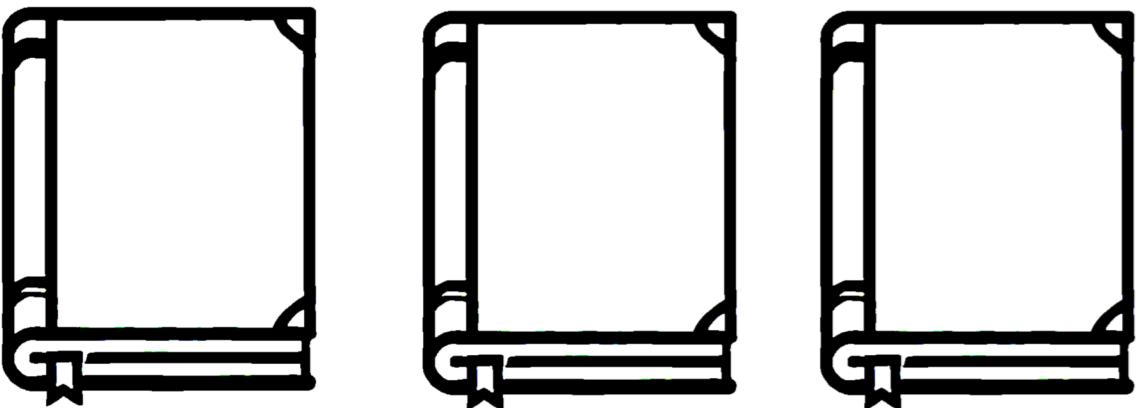
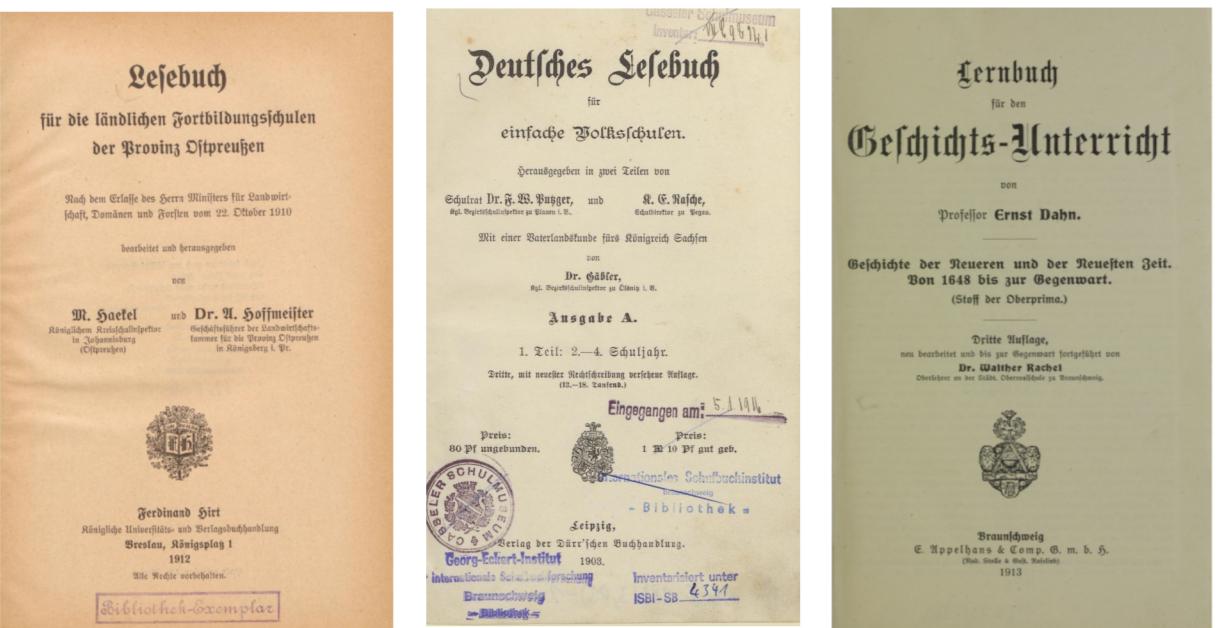


Schulbuchforschung an der Schnittstelle zur Infrastruktur Netzwerkmodellierung mit NER und NEL



Schulbuch

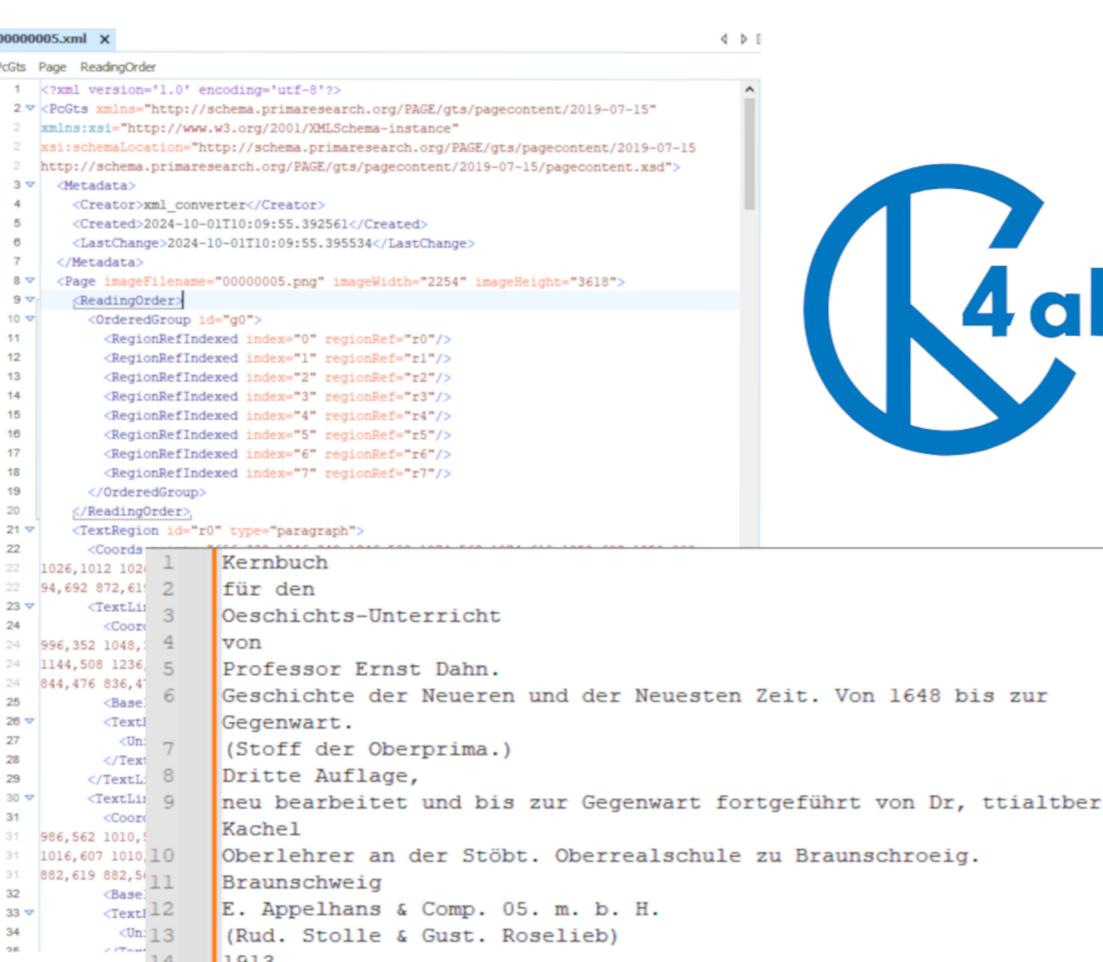


Schulbücher kommunizieren einen Wissenskanon, den eine Gesellschaft als grundlegend erachtet. Der historischen Forschung geben sie jedoch nicht nur Einblicke in die vermittelten Inhalte, sondern auch in die Sinnwelten und Werte einer Zeit. Mit der Einführung der Schulpflicht im langen 19. Jahrhundert erlangte das Schulbuch (bereits vorher ein zentrales Massenmedium europäischer Gesellschaften) eine prominentere Rolle. Es verbreitete die Geschichtsbilder der entstehenden Nationalismen, stiftete Identität und stabilisierte die sozialen Zusammenhänge im Staat. Dieses Ziel erreichte es aber nicht allein durch seine Inhalte, sondern darüber hinaus durch deren zeitgenössische pädagogische Aufbereitung und flächendeckende Verbreitung. Zu diesem Zweck wurden in Deutschland zwischen 1800 und 1945 nach aktuellem Stand etwa 2.200 Geschichtsschulbücher veröffentlicht.

Trotz dieser bedeutenden Rolle bleibt vieles über die Entstehung und Produktion von Schulbüchern unklar. Die Forschung neigt dazu, sich hauptsächlich auf die Analyse der Inhalte zu konzentrieren und dabei die Netzwerke der beteiligten Personen, Verlage und staatlichen Institutionen zu vernachlässigen. Fragen wie „Wer schrieb die Schulbücher?“, „Wer stellte sie her und verbreitete sie?“ und „Welche Netzwerke waren daran beteiligt?“ wurden bisher nur für Einzelbeispiele beantwortet.

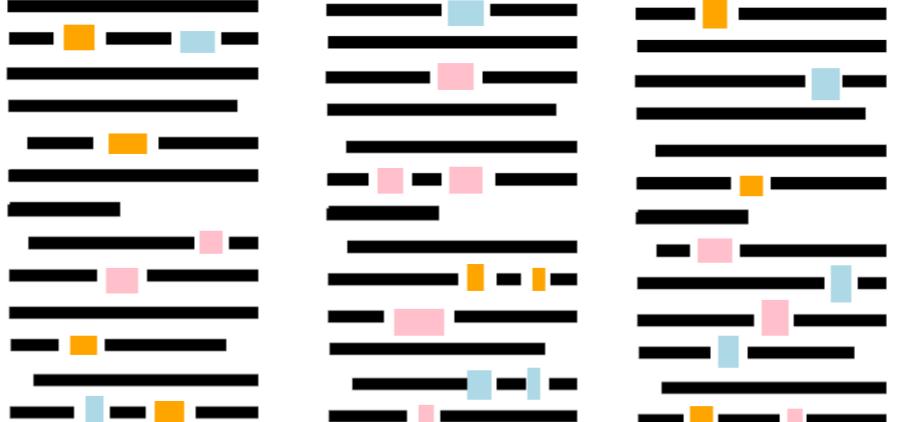


OCR Volltext



OCR-Volltexte bilden die Grundlage von Informationsextraktion mittels NLP-Methoden, zu denen Named Entity Recognition (NER) gehört. Aktuell werden die Schulbuch-Digitalisate und -Volltexte aus dem Kaiserreich durch das Zentrum für Philologie und Digitalität (ZPD) der Universität Würzburg aufbereitet. Im Vordergrund steht dabei die Extraktion von maschinenverarbeitbaren Volltexten aus den Digitalisaten mittels Verfahren der automatischen Texterkennung. Dabei kommen ausschließlich Open Source Werkzeuge, hauptsächlich OCR4all in Kombination mit OCR-D, zum Einsatz. Die primären Datenquellen für diese Untersuchung umfassen bibliographische Datensätze und die digitalisierte Geschichtsschulbücher aus der Zeit von 1871 bis 1918. Die Daten entstanden am Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut (GEI).

Nicht zu verschweigen ist bei dem Projekt die Herausforderung der Arbeit mit den älteren OCRs, die aktuell im Datenkorpus GEI-Digital vorliegen. Diese stammen oft noch aus älteren Projektphasen und spiegeln den Stand der damaligen Techniken und Infrastrukturpraktiken. Daher ist oft eine Neuverarbeitung der Digitalisate notwendig.

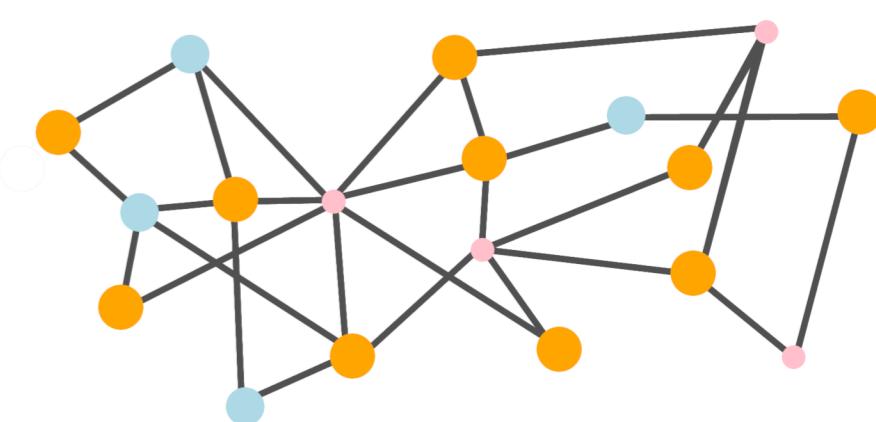


NER [Named Entity Recogniton]

PER	→ GND URI	→ Additional Data
LOC	→ GND URI	→ Additional Data
ORG	→ GND URI	→ Additional Data
PER	→ GND URI	→ Additional Data
LOC	→ GND URI	→ Additional Data
ORG	→ GND URI	→ Additional Data
PER	→ GND URI	→ Additional Data
ORG	→ GND URI	→ Additional Data
LOC	→ GND URI	→ Additional Data

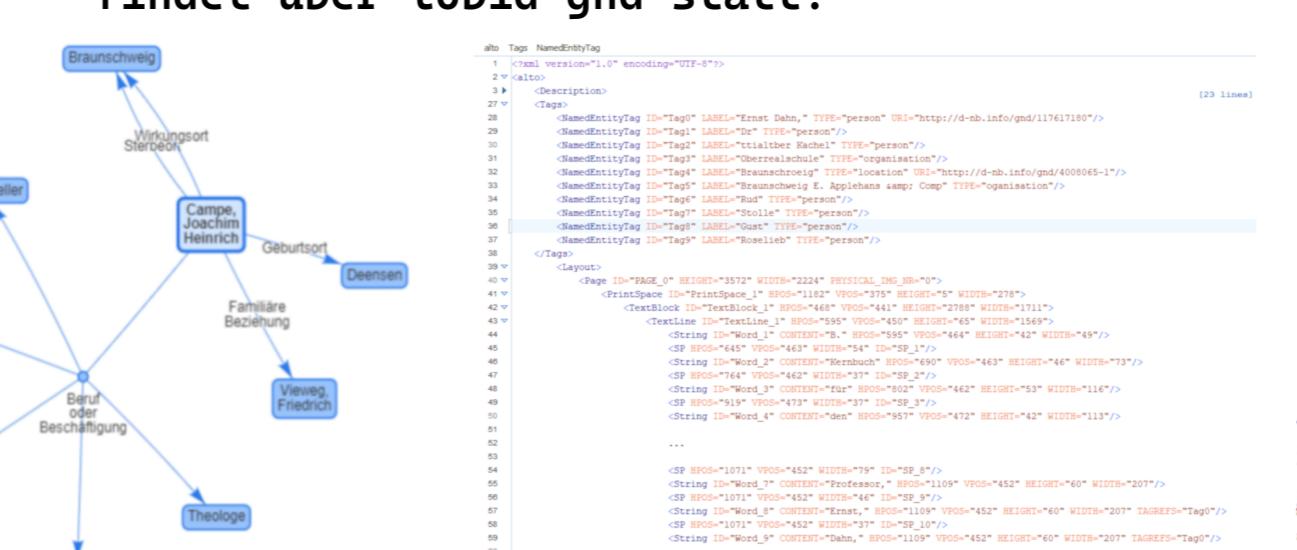
NEL [Named Entity Linking]

[Named Entity Linking]

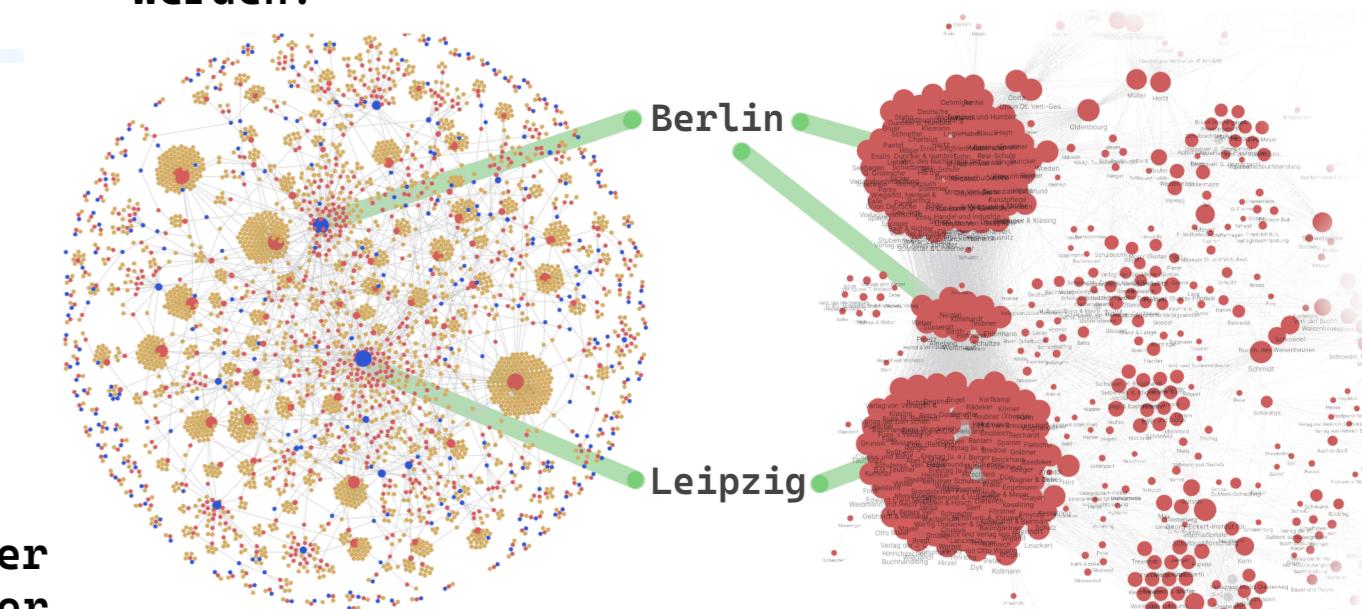


Netzwerk

Die beiden eingesetzten Technologien NER und NEL helfen, die bibliographischen Personendaten zu erweitern und anzureichern. Darüber hinaus können diese Daten aus anderen Projekten am GEI erweitert werden. So z. B. durch die Text Reuse Detection des Projektes SchulbuchEvolution. Das Ergebnis lässt sich in einem neuen Knowledge Graph für die Domäne der Bildungsgeschichte abbilden. Solche Daten sind für eine Netzwerkanalyse natürlich erstmal zu heterogen. Dennoch lassen sie sich in einen für solche Ansätze geeignetes Netzwerk umwandeln, indem bestimmte Typen oder Beziehungen ausgewählt oder abstrahiert werden.



Um die Nachnutzbarkeit der mit NER extrahierten Entitäten zu gewährleisten, werden im anschließenden Named Entity Linking (NEL) zugehörige Normdaten identifiziert. Maßgeblich ist hier Gemeinsamen Normdatei (GND), die von der Deutschen Nationalbibliothek bereitgestellt wird. Obwohl wir auch andere Normdateien wie WikiData in Betracht ziehen, scheint die GND im Moment die beste Option zu sein, insbesondere aufgrund ihrer Integration von bibliographischen Daten. Die Abfrage findet über lobicid-gnd statt.



Mit den bibliographischen Daten allein lässt sich z.B. schon erkennen, wie viele Autoren sich in bestimmten Verlagen und an bestimmten Orten tummeln. Nach der durch NEL möglichen Anreichungen wird aber mehr möglich: U.a. die Frage, wie viele Autoren an derselben Schule gearbeitet haben oder welche Beamten in denselben Ausschüssen saßen, die Schulbücher zur Veröffentlichung genehmigten. Netzwerkdenken wird hier zum Vorteil.

- Literatur**
- Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP.“ *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. <https://doi.org/10.18653/v1/n19-4010>.
 - Ahnert, Ruth, Sebastian E. Ahnert, Catherine Nicole Coleman, and Scott B. Weingart. 2021. *The Network Turn. Changing Perspectives in the Humanities*. Elements in Publishing in the Humanities. Cambridge: Cambridge University Press.
 - Menzel, Sina, Hannes Schnaitter, Josefina Zinck, Vivien Petras, Clemens Neudecker, Kai Labusch, Elena Leitner, and Georg Rehm. 2021. „Named Entity Linking mit Inhaltserweiterung“, herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, und Hans Schürmann, 229–58. Bibliotheks- und Informationspraxis 70. Berlin: De Gruyter.
 - Nöth, Maximilian, Herbert Baier and Christian Reul. 2024. OCR4all 1.0 – Flexible open-source OCR/HTR based on various single-step Solutions. 16th IAPR International Workshop On Document Analysis Systems (DAS) (September).
 - Otto, Marcus. 2018. „Textbook Authors, Authorship, and Author Function“. In *The Palgrave Handbook of Textbook Studies*, herausgegeben von Eckhardt Fuchs und Annekatrin Bock, 95–102. New York: Palgrave Macmillan.