Informationszentrum Lebenswissenschaften
ZB MED

Ein Workshop von:
- TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften
- ZB MED – Informationszentrum Lebenswissenschaften
- ZBW – Leibniz-Informationszentrum Wirtschaft
- Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Workshop Retrodigitalisierung | Abstract

Semantische Anreicherung von Digitalisaten

Nils Heun (Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung)

Bei der Präsentation von Digitalisaten stehen mittlerweile immer stärker die in den Texten enthaltenen Inhalte im Vordergrund. Die Grundlage für deren Erschließung bildet dabei eine erkennungsstarke Optical Character Recognition (OCR), denn sie ermöglicht eine tiefergehende Durchdringung der Texte mittels Named Entity Recognition (NER). Mit ihr können Entitäten, wie Personen, Orte oder Themen in den Texten erkannt und ausgezeichnet werden. Werden diese Entitäten mit Normdaten verlinkt, wie der Gemeinsame(n) Normdatei (GND) oder Wikidata, spricht man von Named Entity Linking (NEL).

Das Archiv der sozialen Demokratie hat die Einführung einer neuen Präsentationssoftware zum Anlass genommen, einen bereits digitalisierten Zeitungskorpus mit insgesamt knapp 300.000 Seiten semantisch aufzuwerten. Im Korpus wurden automatisiert 80.000 Entitäten ausgezeichnet und zur GND und Wikidata verlinkt. Damit stehen Nutzer*innen der Präsentationssoftware vielfältige Optionen zur Recherche zur Verfügung. Neben der klassischen Volltextsuche können sie diverse Indizes durchsuchen, die zusätzliche Filterfunktionen bieten. Damit werden die Rechercheoptionen stark verbessert, da ganz gezielt nach Entitäten gesucht werden kann. Es wird aber nicht nur die Recherche verbessert, sondern auch ein explorativer Einstieg in den Korpus ermöglicht.

Die Suchergebnisse werden in einem Ausschnitt des Digitalisats angezeigt und gehighlightet, womit auf den ersten Blick auch direkt der Kontext sichtbar wird. Es kann aber nicht nur im Digitalisat gelesen werden, sondern auch im zuschaltbaren Volltext, in dem alle ausgezeichneten Entitäten verlinkt sind. Beide Optionen können auch simultan angezeigt werden.

In meinem Vortrag möchte ich unsere gewonnenen Erkenntnisse und die Ergebnisse unserer Arbeit vorstellen. Dabei möchte ich deutlich machen, welche Technologien und Arbeitsschritte notwendig sind, um Digitalisate zeitgemäß präsentieren zu können. Anhand unseres Portals möchte ich zudem vorstellen, wie sich diese für unsere Nutzer*innen darstellen.

Sie finden die digitale Sammlung unter diesem Link:
https://collections.fes.de/historische-presse