E-Science-Tage 2019: Data to Knowledge
27. bis 29. März 2019 in Heidelberg

Details zum Beitrag

134
Art des Beitrags / Konferenztrack: Einreichung eines Abstracts
Format: Kein Format

Transforming data silos into knowledge- Early Chinese Periodicals Online (ECPO)

Matthias Arnold✉, Lena Hessel✉
Organisation(en): Universität Heidelberg, Deutschland

eingereicht von: Matthias Arnold (Universität Heidelberg, DE), ID: 1085
Themen: Fachspezifische oder disziplinübergreifende Herausforderungen, Open Data, Open Science, Open Access

Zusammenfassung

This paper introduces the project “Early Chinese Periodicals Online (ECPO)” [1]. ECPO joins several important digital collections of the early Chinese press and puts them into a single overarching framework. In a first phase, several databases on early women’s periodicals and entertainment publishing were created: “Chinese Women’s Magazines in the Late Qing and Early Republican Period” (*WoMag*), “Chinese Entertainment Newspapers” (*Xiaobao*), and databases hosted at the Academia Sinica in Taiwan. These systems approach the material in two ways: in the *intensive approach* all articles, images, advertisements, and related agents are recorded and assigned to a complete set of scanned pages, while in the *extensive approach* the main characteristic features of publications are stored.

ECPO has begun to join these various materials in a second, ongoing phase of the project. Today, ECPO provides open access to 267 publications comprising over 280.000 pages of print. A key aspect is to make entire issues available, front-to-back, including illustrations, advertisements, and even blank pages. For 138 publications we also provide descriptions of individual items in Chinese with Pinyin transcription. These records also contain genre and column information, basic content analysis, as well as names and roles of agents associated with an item.

Our new cross-database agent service allows us to manage the approximately 47.000 names recorded in *WoMag* and ECPO: a) merge identical names across databases, b) identify agents and assigning names to them, and c) link agent records to authority data (GND, VIAF, Wikidata). Besides creating a curated list of agents occurring in the publications, we also aim to add missing persons to authority files like the GND.

One crucial aspect ECPO is full text capability. Unfortunately, OCR software cannot be used out-of-the-box, for a number of reasons: document analysis fails to recognize complex newspaper layout, character recognition fails when it faces emphasis marks next to characters, and recognized passages have to be grouped in the right semantic order.

The paper will discuss approaches to further exploring and analyzing the knowledge hidden in these publications, together with efforts to open the collection's data for re-use. We will demonstrate workflows in the Agents service and cross-database record curation. We also present results from a crowdsourced approach to newspaper segmentation to generate segments that can easier be OCRed. In addition, we introduce first ideas to create a module for encoding text in TEI and relate it to the database.