

Gius, E., Meister, J. C., Meister, M., Petris, M., Bruck, C., Jacke, J., Schumacher, M., Gerstorfer, D., Flüh, M., and Horstmann, J. (2018-2021). CATMA. Concept DOI: [10.5281/zenodo.1470118](https://doi.org/10.5281/zenodo.1470118).

Helling, P., Jung, K., Reiter, N. and Pielström, S. (2020). Interviewleitfaden zur FDM-Bestandsaufnahme im Schwerpunktprogramm Computational Literary Studies. DOI: [10.5281/zenodo.4269639](https://doi.org/10.5281/zenodo.4269639).

Helling, P., Jung, K. and Pielström, S. (2021). Disziplinspezifisches Forschungsdatenmanagement - FDM-Bedarfserfassung in den Computational Literary Studies. *FORGE 2021 Konferenz: Forschungsdaten in den Geisteswissenschaften - Mapping the Landscape - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen (FORGE 2021)*, Cologne. DOI: [10.5281/zenodo.5379629](https://doi.org/10.5281/zenodo.5379629).

Pempe, W. (2012). Geisteswissenschaften. In: Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J. and Ludwig, J. (eds), *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Boizenburg: Verlag Werner Hülsbusch, pp. 137-60.

Pielström, S., Helling, P. and Jung, K. (2021). Zentralprojekt des DFG-Schwerpunktprogramms Computational Literary Studies. *Program General Meeting*, virtuell. DOI: [10.5281/zenodo.5041338](https://doi.org/10.5281/zenodo.5041338).

Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann M. and Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel. DOI: [10.17175/2020_006](https://doi.org/10.17175/2020_006).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data* 3, Article number: 160018. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).

Notes

1. <https://dfg-spp-cls.github.io/> [last request: 24th of November 2021].
2. DataCite Metadata Schema 4.4, <https://schema.datacite.org/meta/kernel-4.4/> [last request: 07th of December 2021].
3. GitHub, <https://github.com> [last request: 07th of December 2021].
4. Regarding annotations, Eckart and Heid (2014) argue for a separation of content-related interoperability and representation format-related interoperability. For the latter we found the projects in the priority program to agree on CATMA (Gius et al. 2018-2021) using its own TEI Export Format.

Building an OCR Pipeline for a Republican Chinese Entertainment Newspaper

Henke, Konstantin

konstantin.henke@pm.me
Heidelberg Centre for Transcultural Studies, Heidelberg University

Arnold, Matthias

arnold@hcts.uni-heidelberg.de
Heidelberg Centre for Transcultural Studies, Heidelberg University

In recent years, the digitisation of newspapers has made a lot of progress, and large national and international initiatives like Trove¹, Chronicling America², Europeana Newspapers³, Impreso⁴, NewsEye⁵, Oceanic Exchanges⁶, OCR-D⁷, Deutsches Zeitungsportal⁸, and Living with Machines⁹ emerged that are building up on and going beyond sheer digitisation, venturing into various areas of content analysis (Oberbichler et al, 2021). Also, the outcomes of these initiatives are usually provided online with open access, and publications increasingly follow the FAIR principles (Wilkinson et al, 2016). However, most of the textual content covered is printed in Latin script languages, and to a large degree the analytical systems rely on linguistic features like word boundaries, digital lexica, or tagged corpora.

Responding to this from an Asian perspective, i.e. looking at materials from regions where non-Latin scripts prevail, the situation is different. In our case we are working

with newspapers from Republican China. Although there are some projects working on historical Chinese newspapers (Stewart et al, 2020), results have so far rarely been published. Other initiatives provide their final results as commercial products. In general, a certain reluctance can be observed when it comes to publishing research methodologies, not to mention the open access sharing of ground truth, test corpora, or trained models (Arnold et al, forthcoming).

In our project we collected periodicals from the Republican era as image scans (Sung et al, 2014) and started OCR experiments to transform them into machine readable full text.

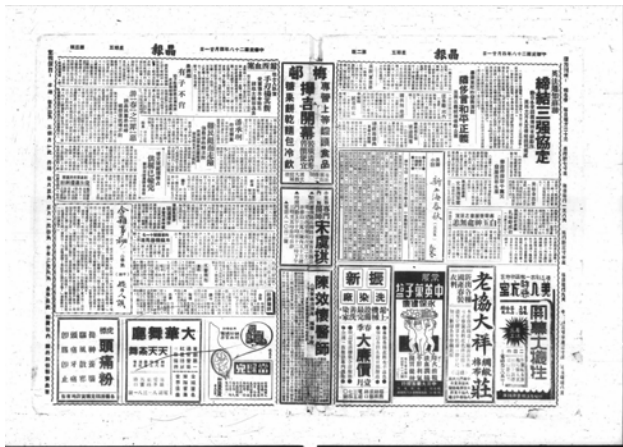


Fig. 1:
One of the 9385 fold scans of *Jing bao*

Looking closer at *Jing bao* 晶報 (*The Crystal*) (cf. Fig. 1), an entertainment newspaper that ran from 1919 to 1940, we soon learned that the key issue of OCR'ing the material actually lies in the page-level segmentation. We therefore started creating ground truth (GT) for geometrical data featuring semantically grouped bounding boxes with labels (article, image, advertisement, marginalia). We then used the resulting dataset to train dhSegment and have the network detect content areas on the folds (Arnold, forthcoming) (Fig. 2).

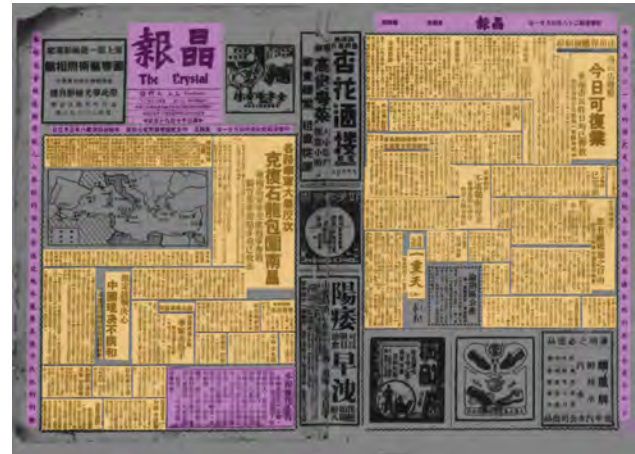


Fig. 2:
Automatic page segmentation results. Blocks with text content are shown in yellow.

Additionally, we created a text GT that not only covers all text in a machine readable local XML format, but also contains information about reading sequence running direction of the text. Based on this GT we were able to process a first set of manual crops, introducing a character segmentation method for grid-based printing layouts which produces over 90,000 labeled images of single characters (Henke, 2021). In this work, a GoogLeNet is trained as an OCR classifier on said character images after extensive pre-training on synthetical character image data created from font files. Additional error correction using language models yields an accuracy of 97.44%.

In our presentation we introduce our work on developing a document image processing pipeline currently focusing on Republican Chinese newspapers with complex layouts like the *Jing bao*. We will present the following concrete contributions:

1. A page-level segmentation approach (as seen in Fig. 2) yielding single text blocks.
2. An OCR pipeline taking single text blocks as input.

While Arnold (forthcoming) presented first promising experiments regarding (1), in this presentation we will concentrate on (2). Our evaluation metric for OCR output is the character error rate (CER) with regard to the ground truth annotation of every text block crop, which, based on the Levenshtein distance, is computed by:

$$CER = \frac{S + D + I}{L}$$

(S, D, I = number of substitutions, deletions, insertions; L = length of the reference sequence, i.e. corresponding GT text).

The character segmentation approach presented in Henke (2021) can however only process text blocks where characters are printed in a grid-like layout, which accounts for a very small portion of the *Jing bao*. Hence, there is a particular need for efficient character detection in less stable layout situations within text blocks, before passing single character images on to the actual OCR engine. As a baseline, we leverage the publicly available state-of-the-art OCR tool Tesseract (Smith, 2007) which provides out-of-the-box segmentation+recognition models even for vertically printed traditional Chinese. Tesseract however seems to struggle with the low input image resolution ($\sim 25 \times 25$ px per character) and overall inconsistent scan quality, leading to a very high CER of **47.85%** on the test set from Henke (2021).

To solve this issue, we use the readily-trained HRCenterNet from Tang et al. (2020) for character detection, and crop the bounding boxes to feed them into the GoogLeNet trained in Henke (2021). However, while our crops have a great variety of aspect ratios, the HRCenterNet expects at least nearly-squared rectangles. Hence, we cut the original images into 250×250 px tiles with a 50 px overlap (both horizontally and vertically, Fig. 3c). Bounding boxes (Fig. 3d) found in the overlapping sections are filtered during the non-maximum suppression (NMS) operation already included in the HRCenterNet pipeline (Fig. 3e).

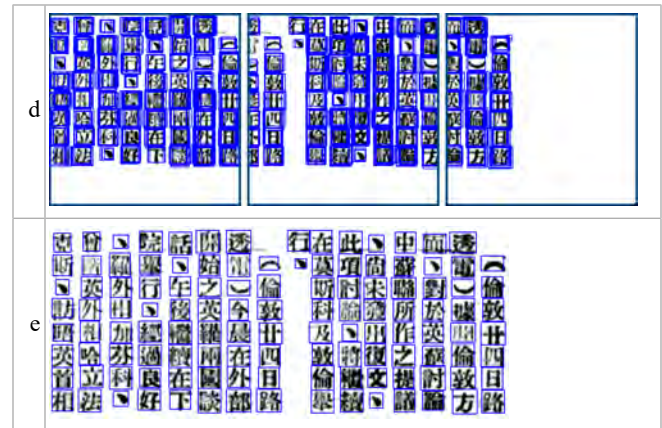


Fig 3:

a) original image, b) image after contrast enhancing, c) tiling with overlap, d) bounding boxes found by HRCenterNet before NMS, e) final result after reconnection of tiles and NMS

In addition, Fig. 4 shows how the HRCenterNet largely profits from contrast-enhancement (Fig 3b) during image pre-processing, especially for low-contrast input images.

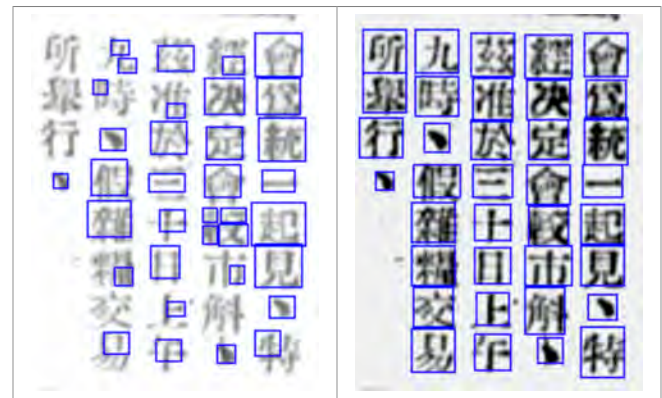


Fig 4:

Effects of contrast enhancement on character detection using HRCenterNet

Using the above method, the CER on the test set of Henke (2021) is reduced to **5.64%**.

In the presentation we will show how the results can be confirmed on a non-grid-based section of the corpus, for which we currently create GT annotations. We are confident that the additional pre-processing of crops and individual character images will help to further reduce the CER, and in combination with (1), yield a powerful document-level OCR pipeline for the *Jing bao* and similar Republican newspapers. This will not only open the door to further processing with the tools of Digital Humanities, but also

further contribute to FAIR-based work in the diverse Asian sphere.

Bibliography

Arnold, M. (2021). Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition. *Digital Studies/Le Champ Numérique*, forthcoming. DOI: 10.11588/heidok.00030918 (preprint).

Arnold, M., Paterson, D. and Xie, J. (forthcoming). Procedural Challenges: Machine Learning Tasks for OCR of Historical CJK Newspapers. *International Journal of Digital Humanities*, Special issue on Digital Humanities and East Asian Studies. (manuscript accepted by special issue editors, currently under review by journal).

Henke, K. (2021). Building and Improving an OCR Classifier for Republican Chinese Newspaper Text. B.A. thesis, Heidelberg University. DOI: 10.11588/heidok.00030845

Oberbichler, S., Boros, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H. and Tolonen, M. (2021). Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians. *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.24565

Smith, Ray (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, pp. 629-633. DOI: 10.1109/ICDAR.2007.4376991

Stewart, S., 朱吟清 Zhu, Y., 吴佩臻 Wu, P., 赵薇 Zhao, W., Gladstone, C., Long, H., Detwyler, A. and So, R. J. (2020). 比较文学研究与数字基础设施建设: 以“民国时期期刊语料库(1918-1949), 基于PhiloLogic4”为例的探索 (Comparative Literature Research and Digital Infrastructure: Taking the ‘Republican China Periodical Corpus (1918-1949), Based on PhiloLogic 4’ as an Example). *数字人文 Digital Humanities*, no. 1: 175–82. online version

Sung, D., Sun, L. and Arnold, M. (2014). The Birth of a Database of Historical Periodicals: Chinese Women’s Magazines in the Late Qing and Early Republican Period. *TSWL* 33, no. 2: 227–37. URL: <https://www.jstor.org/stable/43653333>

Tang, C., Liu, C. and Chiu, P. (2020). HRCenterNet: An Anchorless Approach to Chinese Character Segmentation in Historical Documents. *IEEE International Conference on Big Data (Big Data)*. DOI: 10.1109/BigData50022.2020.9378051

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et

al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. no. 1. *Scientific Data* 3, no. 1: 160018. DOI: 10.1038/sdata.2016.18

Notes

1. <https://trove.nla.gov.au/newspaper/>
2. <https://chroniclingamerica.loc.gov/newspapers/>
3. <http://www.europeana-newspapers.eu/>
4. <https://impresso-project.ch/>
5. <https://www.newseye.eu/>
6. <https://oceanicexchanges.org/>
7. <https://ocr-d.de/en/>
8. <https://www.deutsche-digitale-bibliothek.de/newspaper>
9. <https://livingwithmachines.ac.uk/>

Townsend & Sons, Account Book Manufacturer’s Business Guide and Works Manual: 19th Century Primary Manuscript Source and TEI Encoding

Hermesen, Lisa

lismariehermesen@gmail.com
Rochester Institute of Technology

Walker, Rebekah

rgwtwc@rit.edu
Rochester Institute of Technology

The session will describe the application of TEI guidelines to encode an idiosyncratic primary source manuscript—a volume from the collection of The William Townsend & Sons, Printers, Stationers, and Account Book Manufacturers, Sheffield UK (1830-1910). Volume 3, “Business Guide and Works Manual,” is a remarkable manuscript both for book history and cultural observations about unionization, gender roles, and credit/debit accounting. The extraordinary complexity of the manuscript’s structure requires that it be marked up to render a readable document. This project is using Text Encoding Initiative (TEI) Guidelines to create a digital edition, illuminating such key topics as the Townsend firm’s social networks, information on the men, women and apprentices who worked for the firm between 1830 and 1910, and the web of economic partners with whom the firm did business. The digital edition will be accompanied by documentation of editorial decisions about encoding that