CFP: <https://www.springer.com/journal/42803/updates/18537886>

International Journal of Digital Humanities (Springer) <https://www.springer.com/journal/42803>

Digital humanities constitutes a multilayered umbrella-term encompassing a variety of digitally enhanced projects and engagements from textual analysis to visualization and modeling, network analysis and digitization. The rapid development of this “field” has not affected all areas of the humanities to the same extent. In fact, digital humanities is still focused on the analysis of Western-, particularly English-, language materials, whereas non-Latin script users still face a variety of added challenges in digital research. Recently, a growing demand for “disrupting digital monolingualism” (referring to King’s College London’s recent workshop of the same title) has emerged in conferences and working groups. The peer-reviewed *International Journal of Digital Humanities* wishes to contribute to this endeavor by providing a platform for these underrepresented perspectives in a special issue on digital humanities in the context of East Asian studies.

With this special issue we aim to represent the multifariousness of digital humanities, covering topics from theoretical discussions to applied papers on digital methods in the field of East Asian studies. Potential topics include (but are not limited to): text analysis of East Asian scripts (challenges and solutions), image and visual analysis, GIS, network analysis, topic modeling, OCR and digitization, East Asian digital infrastructure development, digital libraries and archives, software development, curricular developments in East Asian digital humanities, collaborations in digital research. Both individual and co-authored papers may be submitted. Papers submitted to other journals concurrently will not be considered.

**About the Journal**

The International Journal of Digital Humanities is a peer-reviewed academic journal with a focus on digital media and on the development and application of and reflection on digital research methodologies in the humanities. It is concerned with the history, current practice and theory of digital humanities. Further details on this journal are available on the Springer website of [IJDH](https://www.springer.com/journal/42803).

All **proposals** must be written in English and submitted via e-mail to the following address: [ijdh.journal@gmail.com](mailto:ijdh.journal@gmail.com) with the subject: **DH and East Asia**. Proposals for articles must be submitted as an abstract of between 300 and 500 words. In addition to the abstract, the submission must include a reference list (at least 5 items). **All submissions are due by 15 December 2020.**

The program committee will review all submissions and send out notifications of acceptance/rejection by **15 January 2021**. For questions, please, send an email to [jdh.journal@gmail.com](mailto:jdh.journal@gmail.com)

The **articles** should be submitted to the Editorial Manager: [https://www.editorialmanager.com/ijdh/](https://www.editorialmanager.com/ijdh/default.aspx)  by **30 April 2021**.

For **submission guidelines** see the web page of the journal: <https://www.springer.com/journal/42803/submission-guidelines>

Hilde De Weerdt and Aliz Horvath, Guest editors

Gábor Palkó, Editor-in-Chief

**Title**: Procedural Challenges: Machine Learning tasks for OCR of historical CJK newspapers

**Abstract**

The use of convolutional neural networks in digitizing historical documents has drastically expanded the quality and scope of available sources for digital analysis.[[1]](#footnote-2) The ability to reuse or to refine pre-trained models means that DH practitioners are now tackling sources previously deemed impossible to process automatically. In light of these developments, we wish to present our own work on adapting tools developed for use with Western materials to the complex layouts of republican Chinese newspapers from the Early Chinese Periodicals Online (ECPO) project.[[2]](#footnote-3) Our results demonstrate how digital monolinguism negatively affects models and algorithms through cultural biases towards Latinized scripts, and layouts. To overcome these shortcomings, we conducted experiments with crowd-sourcing, pattern recognition, and machine-learning. We aim to provide a development workflow from image scan to machine-readable full-text, and published our ground truths for reuse with other neural networks.[[3]](#footnote-4)

It stands to reason, that similar developments happen in China, and that models trained on CJK sources would be better suited for our task. Yet, we continue to encounter procedural difficulties when following this approach. The first difficulty consists of our ability to locate and identify relevant research. Beginning with the lack of uniformity in search results, for example in China National Knowledge Infrastructure 中国知网 (CNKI)[[4]](#footnote-5) when accessed from China, or from different providers in Europe, including CrossAsia.[[5]](#footnote-6) We are then providing a systematic literature review concerning the use of machine learning methods to process historical CJK documents for either layout-analysis or OCR. Our review includes a discussion of the related problems of a) data-access to ground-truths, and training data, which are preconditions for b) reproducibility and ultimately c) reusability.

In conclusion, we contrast the impressions that our review generates of common research practices in this area in China with the demands placed on publicly funded research in the EU.[[6]](#footnote-7) When it comes to adherence with data management plans,[[7]](#footnote-8) and to the creation of sustainable and FAIR[[8]](#footnote-9) research results the main challenges for collaboration with China are indeed procedural, and less the result of the specifics of the Chinese writing system.

**Figures**

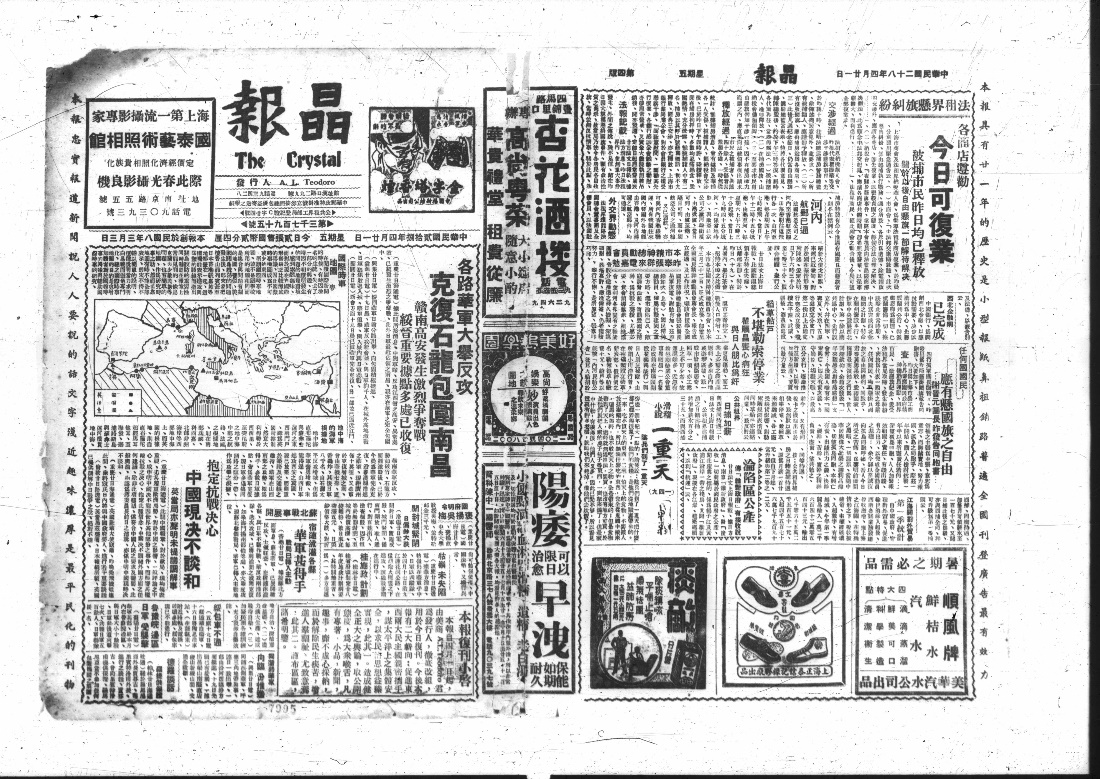


Fig. 1: A typical complex newspaper page layout: Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4. In ECPO: https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/publications.php?magid=1&isid=20&ispage=1.

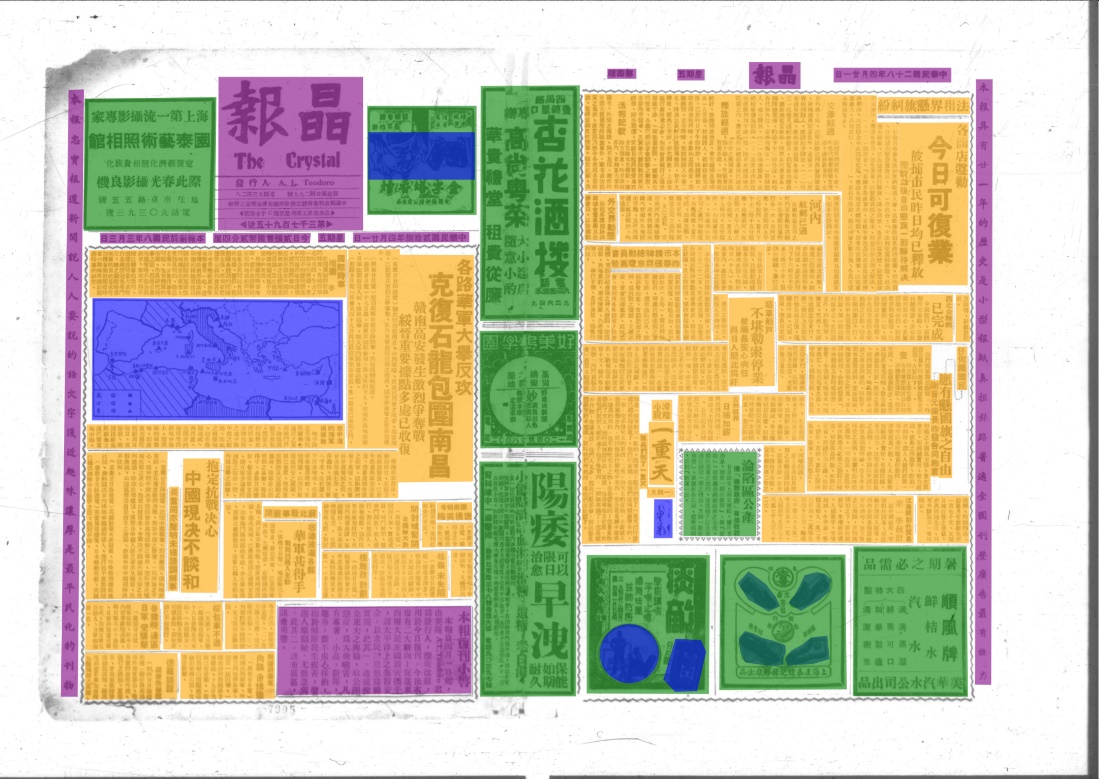


Fig. 2: The page fully manually annotated from our ground truth set (orange = article, blue = image, green = advertisement, purple = header and marginalia). Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4.

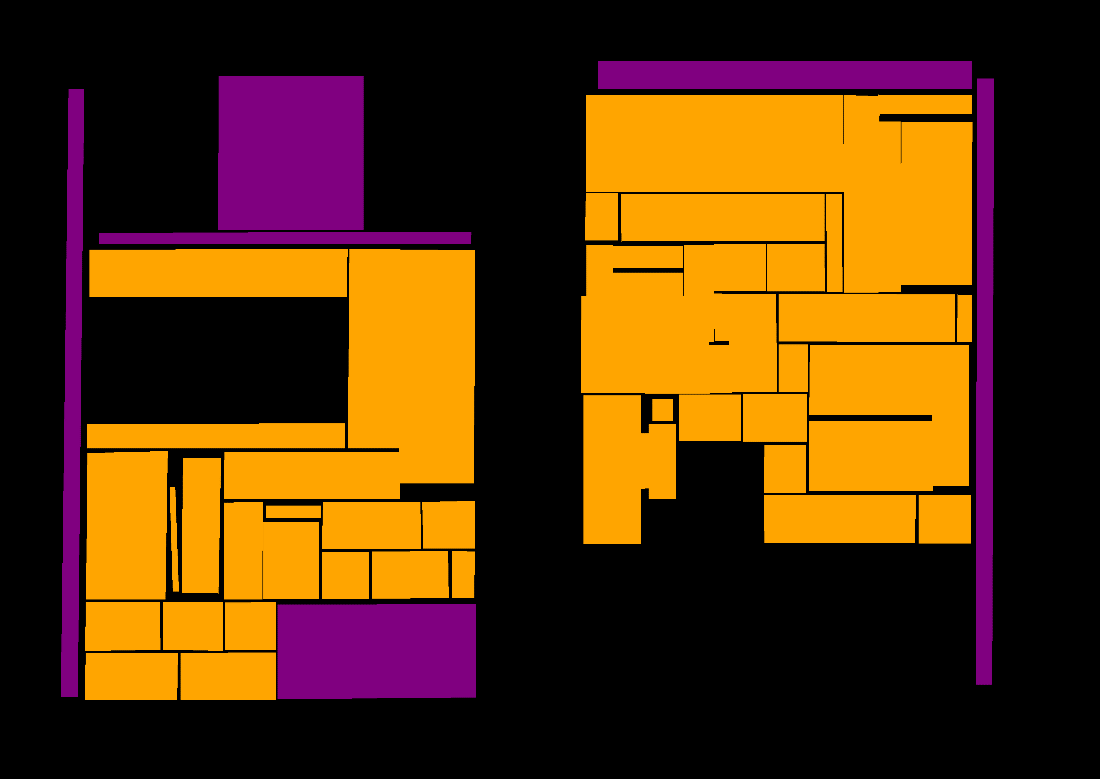


Fig. 3: Detection of content types using dhSegment: advertisements and images ignored, orange = text (i. e. “articles”), purple = header and marginalia. Jing bao 晶報 (The Crystal), April 21, 1939, pages 1+4.

**References** (selection of Chinese research articles):

张玮. “民国报纸数字化验收常见问题研究——以国家图书馆为例.” 图书情报研究, no. 3 (2019): 72–79.

方自金. “民国报纸影印出版的现状、问题与建议.” 出版参考 2019, no. 6 (2019). [www.sohu.com/a/330753909\_488898](https://doi.org/www.sohu.com/a/330753909_488898).

郭薇, and 任思琪. “报纸数字对象模型设计与应用——以上海图书馆《全国报刊索引》平台为例.” 图书馆杂志 2018, no. 7 (2018): 41–52.

杨海林. “基于深度学习的古籍文字识别和检测.” 硕士. 华南理工大学. 2019.

郭利敏, 葛亮, and 刘悦如. “卷积神经网络在古籍汉字识别中的应用实践.” 图书馆论坛 39, no. 10 (2019): 142–48.

**Bibliography**

Ares Oliveira, Sofia, Benoit Seguin, and Frederic Kaplan. “DhSegment: A Generic Deep-Learning Approach for Document Segmentation.” In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12. IEEE, 2018. https://doi.org/10.1109/ICFHR-2018.2018.00011.

Arnold, Matthias. “Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition.” *Digital Studies / Le Champ Numérique* 11 (2021).

Arnold, Matthias, and Lena Hessel. “Transforming Data Silos into Knowledge: Early Chinese Periodicals Online (ECPO).” In *Heuveline, Vincent , Gebhart, Fabian Und Mohammadianbisheh, Nina (Hrsg.): E-Science-Tage 2019: Data to Knowledge*, 95–109. Heidelberg: heiBOOKS, 2020. https://doi.org/10.11588/heibooks.598.c8420.

European Research Council (ERC). “Guidelines on Implementation of Open Access to Scientific Publications and Research Data,” April 21, 2017. https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide\_en.pdf.

Liebl, Bernhard, and Manuel Burghardt. “An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers,” April 15, 2020. https://arxiv.org/abs/2004.07317v1.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3, no. 1 (March 15, 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Researchers in Europe are following quite strict guidelines in their research:[[9]](#footnote-10) Project applications require data management plans[[10]](#footnote-11), outcome should be sustainable and follow the FAIR guiding principles[[11]](#footnote-12) including open access not only to the papers, but also the data sets, models, code/algorithms, and outcome of experiments. These principles also apply to DH research in East Asian Studies.

[[12]](#footnote-13)We are developing workflows to automatically process newspaper image scans to machine readable full text. To do so, we experimented with crowd-sourcing, pattern recognition, and machine-learning. We created ground truth data and use them with neural networks.[[13]](#footnote-14) To situate our approaches within the landscape of ongoing research we attempted to get an overview of the state of the field. We wanted to find out about good (if not “best”) practice, perhaps find papers discussing existing data models, look into their data sets to possibly re-use them, or compare their results to adopt and possibly improve them.

Our hypothesis is that – similar to DH studies in the western world - there are numerous projects applying OCR technologies and-more recently-neural networks. We present our investigations in Chinese language research papers and illustrate some of the hurdles and pitfalls we encountered.

The paper gives an overview of our findings. We faced a number of challenges during this process. One aspect is accessibility: this begins with simple exercises like searching for relevant publications. Where is current DH research from East Asia published? How and by whom can it be accessed? This may sound trivial, but although large academic research platforms like the China National Knowledge Infrastructure 中国知网 (CNKI)[[14]](#footnote-15) can be used within Europe, not all providers offer all CNKI data sets, even if they are provided by national services like CrossAsia[[15]](#footnote-16) in Germany.

The next aspects are reproducibility and re-usability: if papers are accessible, is this also true for the systems and platforms the authors introduce? If researchers introduced data sets or data models - are they available to reproduce the results or re-use the methodology? Even if these rather technical issues are solved and papers, data sets or models can be re-used, there still may exist some structural pitfalls and implications in relation to the inter-continental data sharing.

Schlussparagraph

(411 Wörter - ohne anmerkungen)

Should we explicitly cover Japanese and Korean OCR?

1. Bernhard Liebl and Manuel Burghardt, “An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers,” April 15, 2020, <https://arxiv.org>.   
   Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan, “DhSegment: A Generic Deep-Learning Approach for Document Segmentation,” in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (IEEE, 2018), 7–12. [↑](#footnote-ref-2)
2. Early Chinese Periodicals Online (ECPO), <https://uni-heidelberg.de/ecpo>. [↑](#footnote-ref-3)
3. Matthias Arnold and Lena Hessel, “Transforming Data Silos into Knowledge: Early Chinese Periodicals Online (ECPO),” in *Heuveline, Vincent , Gebhart, Fabian Und Mohammadianbisheh, Nina (Hrsg.): E-Science-Tage 2019: Data to Knowledge* (Heidelberg: heiBOOKS, 2020), 95–109, https://doi.org/10.11588/heibooks.598.c8420; Matthias Arnold, “Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition,” *Digital Studies / Le Champ Numérique* 11 (forthcoming). [↑](#footnote-ref-4)
4. Chinese version: <https://cnki.net/> ; overseas version: <https://oversea.cnki.net/index/> [↑](#footnote-ref-5)
5. CrossAsia <https://blog.crossasia.org/about/?lang=en> is one of the portals of the Specialized Information Service Asia (FID Asia), funded by the German Research Foundation (DFG), cf. <https://idw-online.de/en/news646187> (in German). [↑](#footnote-ref-6)
6. European Research Council (ERC), “Guidelines on Implementation of Open Access to Scientific Publications and Research Data,” April 21, 2017, https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide\_en.pdf. [↑](#footnote-ref-7)
7. Many institutions provide examples and guides, for a basic introduction see https://en.wikipedia.org/wiki/Data\_management\_plan. In Germany, two initiatives provide more information: https://forschungsdaten.org - co-funded by the German Research Foundation (DFG), and https://forschungsdaten.info - funded by the State of Baden-Württemberg [↑](#footnote-ref-8)
8. Mark D. Wilkinson et al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship,” *Scientific Data* 3, no. 1 (March 15, 2016): 160018, https://doi.org/10.1038/sdata.2016.18. [↑](#footnote-ref-9)
9. European Research Council (ERC), “Guidelines on Implementation of Open Access to Scientific Publications and Research Data.” [↑](#footnote-ref-10)
10. Many institutions provide examples and guides, for a basic introduction see https://en.wikipedia.org/wiki/Data\_management\_plan. In Germany, two initiatives provide more information: https://forschungsdaten.org - co-funded by the German Research Foundation (DFG), and https://forschungsdaten.info - funded by the State of Baden-Württemberg [↑](#footnote-ref-11)
11. Wilkinson et al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” [↑](#footnote-ref-12)
12. [↑](#footnote-ref-13)
13. Arnold and Hessel, “Transforming Data Silos into Knowledge”; Arnold, “Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition.” [↑](#footnote-ref-14)
14. Chinese version: <https://cnki.net/> ; overseas version: <https://oversea.cnki.net/index/> [↑](#footnote-ref-15)
15. CrossAsia <https://blog.crossasia.org/about/?lang=en> is one of the portals of the Specialized Information Service Asia (FID Asia), funded by the German Research Foundation (DFG), cf. <https://idw-online.de/en/news646187> (in German). [↑](#footnote-ref-16)