Articles on proofing and verification of indexing newspaper contents

(Zhang Wei 2019; Xiao Hong and Huai Yan 2017; Xiao Hong 2017)

(Qian Chengjun 2008)

In 2014, the National Library launched the digitization project of Republican China newspapers. They decided that the project uses the image scans of each page as basis. They used OCR to recognize item 篇目 (article) titles, and marked the item’s position of each on the page, to use both for the title search.

Beginning with MF digitization in 1990s(?)

Provided by library or also digitized by external agencies, or did they digitize from MF rolls?

Outsourcing to structure scans, edit newspaper and issue metadata, and analyse content by page: draw boxes around columns, titles, authors, texts, identify parts of an article, record text

Text “OCR” - > manually or with software?

Staff for work mostly “junior high school and and high school students”

Problem of libraries: data output with many errors, outsourcing dominant way to digitize newspapers, many different libraries do it, therefore the articles as guideline

Different levels of errors

Suggestions for improvement (but automatical approaches e.g. to fulltext not mentioned)

Qian Chengjun, 钱承军. 2008. “College Libraries Should Pay Attention to the Conservation and Utilization of Republican Literature”, 高校图书馆应重视对民国文献的保护与利用 (Gaoxiao tushuguan yin zhongshi dui minguo wenxian de baohu yu liyong). *Shandong Library Quarterly*, 山东图书馆季刊 (Shandong tushuguan jikan), no. 1: 74–77.

Xiao Hong, 肖红. 2017. “The Main Problems and Solutions in the Practice of Digitization of Newspapers in the Republic of China”, 民国报纸数字化实践中的主要问题及处理策略 (Minguo baoyhi shuzihua shijian zhong de zhuyao wenti ji chuli celüe). *Researches on Library Science*, 图书馆学研究 (Tushuguan xue yanjiu) 2017 (4): 22–37. https://doi.org/10.15941/j.cnki.issn1001-0424.2017.04.004.

Xiao Hong, 肖红 and Huai Yan, 槐燕. 2017. “An Analysis of Quality Checking Problems in the Practice of Digitization of Newspapers of Republic of China”, 民国报纸数字化实践中的质检问题探析 (Minguo baozhi shuzihua shijian zhong de zhijian wenti tanxi). *Researches on Library Science*, 图书馆学研究 (Tushuguan xue yanjiu), no. 07: 61-78+87.

Zhang Wei, 张玮. 2019. “Research on the Common Problems in Digital Acceptance of Newspapers in the Period of Republic of China: A Case Study of National Library of China”, 民国报纸数字化验收常见问题研究——以国家图书馆为例 (Minguo baozhi shuzihua yanshou changjian wenti yanjiu——yi guojia tushuguan wei li). *Library and Information Studies*, 图书情报研究 (Tushu qingbao yanjiu), no. 3: 72–79.