(Xiao 2017)

Xiao, Hong [肖红]. 2017. “民国报纸数字化实践中的主要问题及处理策略” [The Main Problems and Solutions in the Practice of Digitization of Newspapers in the Republic of China]. 图书馆学研究 [Researches on Library Science] 2017 (4): 22–37. https://doi.org/10.15941/j.cnki.issn1001-0424.2017.04.004.

Digitization of republican newspapers has three types: 1) scan only into electronic images 扫描成电子图像; 2) article title OCR 文章标题OC; 3) full text recognition 全文识别. The different types allow different services: 1) only browsing by publication title or page, no search is possible, production costs are low. 2) allows title search but can only locate the page where the title can be found, the specific location of the title has to be found by the reader, production costs relatively low. 3) allows searching based on full text, has comprehensive information retrieval, at relatively high cost.

newspaper content, layout utilization, article content and layout diversification 报纸内容、版面利用、文章篇目内容及排版多样化, deeply affect the digitization

2014 National Library launched digitization of republican newspapers, and decided to perform article title OCR. Until 2015 they digitized 21 newspapers种, with 368786 pages版 and 3864562 items篇 (cf. Table 1). The author was involved for 3 years, recorded and analysed special problems 特殊问题, through praxis gained experience about how to handle these issues 处理经验.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 记录标识号 | 报名 | 总期数 | 版面数 | 篇目数 |
| 1 | 00N001037 | 《大刚报》 | 2.097 | 10.173 | 137.862 |
| 2 | 00N000846 | 《新华日报》 | 3.231 | 12.447 | 133.068 |
| 3 | 00N000203 | 《益世报》 | 8.847 | 72.488 | 719.078 |
| 4 | 00N001043 | 《华北日报》 | 3.934 | 34.772 | 396.060 |
| 5 | 00N001044 | 《中央日报—福建》 | 1.405 | 7.666 | 125.193 |
| 6 | 00N001720 | 《中央日报—昆明》 | 3.724 | 17.580 | 217.725 |
| 7 | 00N001745 | 《中央日报—贵阳》 | 3.214 | 14.597 | 156.759 |
| 8 | 00N001428 | 《阵中日报》 | 3.426 | 7.972 | 123.015 |
| 9 | 00N001445 | 《福建日报》 | 1.962 | 5.124 | 73.076 |
| 10 | 00N001724 | 《冀中日报》 | 841 | 3.204 | 25.078 |
| 11 | 00N002542 | 《解放日报》 | 2.130 | 7.992 | 92.647 |
| 12 | 00N001021 | 《西北文化日报》 | 5.865 | 24.939 | 286.164 |
| 13 | 00N001024 | 《甘肃民国日报》 | 4.933 | 20.498 | 237.844 |
| 14 | 00N001453 | 《南宁民国日报》 | 2.567 | 17.507 | 180.695 |
| 15 | 00N001755 | 《革命日报》 | 3.578 | 14.530 | 167.925 |
| 16 | 00N000638 | 《山东民国日报》 | 784 | 6.164 | 50.841 |
| 17 | 00N000470 | 《新天津》 | 4.689 | 53.946 | 422.946 |
| 18 | 00N001721 | 《青岛日报》 | 1.891 | 21.793 | 138.570 |
| 19 | 00N001694 | 《国风日报》 | 2.010 | 6.234 | 65.128 |
| 20 | 00N000474 | 《工商日报》 | 2.036 | 5.282 | 67.437 |
| 21 | 00N001674 | 《贵州商报》 | 969 | 3.878 | 47.451 |
|  | 合计 |  | 64.133 | 368.786 | 3.864.562 |

Scanning method

Grey-scale vs binary, comparison (Table 2), images 3x more storage, but good when text and images mixed, therefore **grayscale scanning** chosen

Based on Image Data Processing Standards and Operation Guidelines 《国家图书馆图像数据加工标准和操作指南》chose 300dpi

Naming images 001, 002, by page, if multiple pages: duplicate image and name 001, 002

1.2 OCR

1.2.1 Hierarchical structure 层级结构

Each issue期 contains multiple pages版, with 4 and 8 being the majority; a few newspapers have 2 or 12, or even more. Each page 版 of the newspaper contains multiple columns 栏目or directly separate articles篇目文章, under the columns are multiple separate articles in the same column, and in a few cases there are small columns appearing under the columns. Each article 文章contains title标题, body正文, author作者, and attached pictures附属图片. (Figure 4)

1.2.2 XML text

Intro XML, because newspaper structure ultimately use XML, each page is one XML file

1. 3 Quality inspection of XML files

Importance of accuracy of volume卷期, page版次, title 标题and author作者 data, especially title

Example use of wrong characters: "Mao Zedong"毛泽东, incorrectly identified as "Mao Yidong"毛译东

Location of article also important, since 12-30 articles on one page are possible and usually one article is divided into multiple small areas, sometimes even within the area of one article can be other information. The boxes were labelled with Arabic numbers, but since this is hard to check this, the team developed a tool to verify xml data with the image (Figure 6)

In the tool, the xml structure of one page is shown together with all numbered articles in one area, in the second are the metadata is shown, and in the third area the coordinates of all article parts are shown on the image scan. During the verification the respective boxes are highlighted with yellow background with blue and red outlines, based on the coordinates

Unfortunately, the colors are undiscernible in the black and white figure

1. 4 Special problems in practice

600 publications, 3.15 mio pages, 30 mio items from republican newspapers

besides good preparation, some special issues occur or problems were not foreseen during planning

occur on reporting level报种层面, page level版面层面, or item level篇目级别问题

2. Reporting issues

Changes in publication frequency and number of editions are common, no problem

newspaper renamings 报纸更名, changes of publishing houses and publishing locations 出版社和出版地的变化 require special treatment

2. 1 newspaper name change  
dates are not continuous and the intervals between adjacent publication periods are irregular.

Examples: As shown in Figure 8, the "Revolutionary Daily" 《革命日报》was first published in 1935, and before January 1, 1940, all applications were "Revolutionary Daily"《革命日报》. On January 1, 1940, it was renamed "Guizhou Daily"《贵州日报》.  
As shown in Figure 9, Yishibao《益世报》 has been renamed many times since it was founded in Tianjin in 1915. It was Yishibao before October 31, 1937, and it was Yishibao from November 1, 1937 to May 5, 1939. Beijing Yishibao《北京益世报》, which was renamed as “Yishibao Beiping” 《益世报北平》on August 15, 1946.  
As shown in Figure 10, "New Tianjin"《新天津》 was first published on September 10, 1924, and was renamed "New Tianjin News" 《新天津报》on May 1, 1938.  
The predecessor of "Central Daily" (Guiyang edition) 《中央日报》( 贵阳版) was "Wuhan Daily"《武汉日报》, and the predecessor of "Central Daily" (Fujian edition) 《中央日报》( 福建版)was "Fujian Minbao"《福建民报》.

2. 2 Newspaper office relocation  
Examples: 《中央日报》《大刚报》《新华日报》

2. 3 processing strategy

Users cannot understand that 《贵州日报》and《革命日报》are identical. Search by name not enough.

Newspaper XML metadata has to be expanded by this information, incl. remarks备注cf. Fig. 11: example with 题名备注title remarks. If title and title remarks are both searched, both titels can be found, similar: place of publication and publishing house

* Publishing information in ECPO important, alternative titles or other controlled lists should be used instead of strings in remarks field

3 Page/issue problems版面问题

Republican period: no rules and standards in typesetting and printing 排版印刷. Page sizes vary, portrait or landscape format not fixed, reading directions vary

Often regional newspapers produce one newspaper together, so its common to see special publications, such as supplements增刊, special issues特刊, supplements副刊, and pictorials画刊

3. 1 Newspaper subsidiary publication  
commonly seen: supplements增刊, special issues特刊, supplements 副刊and extras号外. These additional publications are within the regular newspaper, e.g p.1-4 , and 5-8 supplements. In other cases subsidiary publications are named separately as independent entities

Examples: 《益世白话报》、《女子周刊》为《益世报》的副刊"Yishe Vernacular Newspaper" and "Women's Weekly" are supplements of "Yishe Newspaper". “Yishe Vernacular Newspaper” shows four smaller paegs on one sheet, with upper and lower pages in opposite reading direction.

《济南市政周刊》为《山东民国日报》的副刊

Fig. 12-16

3. 2 Multiple Newspapers Joint Edition

Example Fig.17

3. 3 combined issues两版互通  
"universal edition"“通版” in "New Tianjin" 《新天津》published on February 22, 1939

3. 4 text direction  
see example above, of “Yishe Vernacular Newspaper”

3. 5 processing strategy  
3. 5. 1 for affiliated publications

involves the naming of the image files of the subsidiary journal and the embodiment of the name of the subsidiary journal

If the printed editions of the subsidiary publications are sorted along the regular pages of the newspaper, they can be named according to the regular editions, need to record the names of the subsidiary publications in the column of each article in the XML

If the print edition of the subsidiary publication is independent of the regular edition, starting from the first edition, the subsidiary publications are classified as extra, supplement, supplement, and special

file naming rules are: Pinyin initials (uppercase) of special layout types + edition number (2 digits, if there is an actual printed edition, name it according to the actual name, if not, name it in the order starting from 01), such as the first edition H01 号外第一版H01, Supplement First Edition Z01增刊第一版Z01, Supplement First Edition F01副刊第一版F01, Special Edition First Edition T01特刊第一版T01.

3. 5. 2 for joint newspapers  
only digitized once, is included in the XML metadata of one single newspaper

3. 5. 3 for combined issues (tongban通版)  
need to be joined in digitization process (stitching)

3. 5. 4 for text direction  
text direction needs to be adjusted to conform to the normal reading habits

4 article problems 篇目问题

Articles 篇目文章are the smallest unit in the digital processing structure data of newspapers in the Republic of China, usually composed of quotes引题, headlines标题, subtopics副题, main text正文, pictures 图片and authors作者.

Content and typesetting 排版of the titles is varied and irregular

typical are nested articles篇目嵌套篇目, various advertising articles各种广告类篇目, white space caused by newspaper clippings剪报导致的留白, various pictures 各类图片and many untitled articles无标题的篇目.

4. 1 nested content

is a closed content area that contains other independent content

Example fig. 20《大刚报》1945 年11月25 日第二版

4.2 Advertising

During the Republic of China, newspapers were considered the cheapest and most effective advertising medium. After the May Fourth Movement, there were almost as many advertisements in a newspaper as news. The role of newspaper advertisements is not only economic, but also a new form of communication and social activities between businessmen and people in the Republic of China, including commercial advertisements and various non-commercial profit advertisements.

Ge Gongzhen 戈公振 discussed ads, cf. 蒋忠波.论戈公振的新闻编辑思想及其现代意义[J]. 编辑之友， 2010（3）: 101-104. And FN1: 赵琛. 民国报纸广告［J］. 中国广告，2005 ( 4) : 64 － 69.  
According to statistics, the advertisements in these newspapers mainly include the following types of content:

* commercial promotion involving 社会广告
  + commercial affairs, commodities, finance, prices, machinery, medicine, and luxury goods; gatherings, laws, recruitment, charity, entertainment, gambling, etc. Advertisement"; 商事、商品、金融、物价、机器、医药、奢侈品的商务推广; 集会、法律、招寻、慈善、游艺、赌博等“社会广告”;
* cultural advertisements 文化广告 on education (referring to school enrollment, school opening, exhibitions, etc.), books, movies, and various publications; 教育( 指学校招生、开学、展览会等) 、书籍、电影及各种出版物
* there are traffic advertisements 交通广告on flight schedules, train schedules, posts and telecommunications航期、车班、邮电
* miscellaneous advertisements杂项广告

within one type of newspaper the layout position of the advertisement on the page is relatively fixed, just content is slightly different

example: 《大刚报》第一版，《京报》第一、第四版are all advertisements, incl. announcements from this newspaper, various stores and product promotions, new book recommendations, and announcements from certain organizations 本报启事、各种店铺和商品宣传、新书推荐及某些单位机构的告示

4. 3 News clippings lead to missing content

cut off by readers before the newspaper’s collection, blank area

difficult to determine which pieces of content are missing (both pages)

4. 4 pictures

4. 5 Untitled articles

4.6 Processing strategy

5 Conclusion

The National Library has more than 600 types of newspapers and more than 3 million editions of the Republic of China. The digital processing of 21 types and 360,000 editions of the Republic of China newspapers has been completed in the two years from 2014 to 2015. **It is expected that a total of 100 types and 600,000 editions of the Republic of China will be completed in February 2017**.

Compared with the entire collection of newspapers in the Republic of China, less than one-fifth has been completed at present, and there is still a long way to go before full digitization.