PDF Converter Production of Historical Newspaper Digitization: the picture experience of China's DaChengLaoJiu Database

DING Xiaowen ; HUANG Weiqun
（Jiangsu Administration Institute ;Shenzhen Administration Institute ）

1 Introduction

The digital application of historical newspapers, means the display of original historical newspapers, articles and pictures on screen via computer and web technology, based on the adequate protection of historical data, which can also sort out more valuable subject materials through data mining technology. A series of exploration and practice around historical newspaper digitization and related issues such as China's DaChengLaoJiu database in 2010，Dazhong Daily historical newspaper digitalization project in 2011, and the digitalization converter production in Beijing Company in 2012, etc. are conducted, including printed newspaper and film scanning，OCR（optical character recognition) and proofreading, metadata extraction and indexing, PDF construction technology, etc.

2. The Survey on Chinese historical newspaper digitization projects

Currently，the companies engaged in Chinese historical newspaper digitization including DaChengLaoJiu, etc., as shown in Table 1:

Table 1: Chinese historical newspaper digitization projects

Survey on Chinese historical newspaper digitization projects SURVEY DATE：2014 -5-26						
Company	Image capturing	OCR	Metadata extraction	Classified indexing	Full-text Database	Cellphone Tablet PC
DaChengLaoJiu	Single layer	×	×	√	√	×
Dazhong Daily	Double layer	√	√	√	√	√
Beijing Company	Double layer	√	√	√	√	√
National newspapers	Single layer	×	×	√	×	×

and periodicals index						
Duxiu platform	Double layer	√	×	√	√	√

However, the key technologies of Chinese historical newspapers digitization are diversified. Some specialize in image capturing by scanning, including print scanning and film scanning, etc. Some specialize in OCR and proofreading, layout analysis and division, making the layout of revised scanned images according to the themes of articles. Some specialize in metadata extraction and classified indexing, making the automatic identification and automatic extraction towards subject, subtitle, introduction, author, source, keywords, abstract, references and the external characteristics of articles.

3. The cases of PDF format files production of Dazhong Daily and Beijing Company

For the early historical newspapers, as there are no corresponding electronic files, so you need to make double layer or refactor the PDF. The focus of double PDF production is as follows: scanning images and processing them into compressed images of appropriate clarity which will be used as the upper image layer of double PDF; rearranging the text according to the original layout structure to form the hidden lower text layer. However, refactoring PDF uses images and text data to make the whole graphic mixed rearrangement according to the original layout structure, which is a single layer structure.

3.1 Working process

3.1.1 Newspapers checking. According to the dates and pages, count the number of newspapers, confirm their completeness, identifiability and scannability. As newspapers generally have a certain circulation (usually more than one copy), the project should choose the best copy as far as possible. If the printed newspapers are not available, the microfilm can substitute.

3.1.2 Image scanning and modification. Use large scanners scan printed newspapers，use film scanners scan microfilm. Modify the scanned TIF images, remove the stains and cracks.

3.1.3 OCR Character recognition and proofreading. OCR is a technology which can recognize the characters on pictures automatically via computer, the OCR recognition accuracy of standard printing Chinese characters can reach more than 99%. Due to the simple technology of early newspapers printing and saving environment, the recognition rate may be slightly lower, thus the scanning need to be proofread many times to ensure the quality. Proofreading includes artificial

proofreading and intelligent automatic proofreading.

3.1.4 Layout analysis and division. Make the layout and identification of revised scanned images according to the themes of articles.

3.1.5 Making format files. Make the searchable digital format files according to steps 2, 3, 4, such as PDF, etc.

3.1.6 Digital data checking. Check the digital data such as words, pictures and file formats from the above steps again, to ensure the completeness and correctness.

3.1.7 Data warehousing. Characters are put into the full-text database, images are put into the image database, and format files are put into the format file database. The above three databases are associated, which have the federated searching function.

3.1.8 To set up double-platform retrieval system. Generally adopts B/S architecture, the users can retrieve the above three databases through the browser.

From the perspective of transformation, the above processes can be divided as follows: Setting transformation standard and data standard→ Digital processing→Layout analysis and attributes annotation → Words segmentation → Identification and proofreading → XML text production→PDF file format reduction and other steps.

3.2 The main differences of two kinds of PDF

(1)on the structure of PDF structure：double PDF production，has two logistic layers(one is image layer and the other text layer).The upper layer is visible images for browsing (in order to control the file size, the picture layer is generally scanned images using high-definition compression format), which can show original scanned pages. The lower layer is a hidden text layer for text retrieval (not visible when browsed). The text in the text layer is the text layout image after correction by OCR, which is the same with that in the upper image layer, though one is a picture of the text and the other is the text of a text. Such double PDF can completely keep the the effect of original layout, and can be selected, copied, and retrieved through the text of the lower layer. Reconstructive PDF is a contemporarily popular single image-text structure.

(2)On PDF format rearrangement: As to historical newspapers, double and reconstructive PDF should be carried out by layout rearrangements. The double layer PDF is rearranged according to the original layout, while PDF reconstruction has to follow the way of today's image-text, which makes the workload heavier. For the recent historical newspapers after digitization of publishing, because of printable PS layout file and the corresponding digital vector font ,which can generate

accurately bulk of today's popular image-text PDF version, there is no need for the production of double or reconstructed PDF.

(3)On Visual browsing: double PDF 100% maintains scanned layout visual effects, though by the limited accuracy of the picture layer, in which the text will be deformed when downsized and produce a mosaic blur when enlarged.On the contrary, characters in the reconstructed PDF text can keep the perfect visionary effect. However, the reconstructed PDF text fonts may be different from the original font. Especially for the early type or mimeographed newspapers, because there is no corresponding figure vector font, it is notable to 100% keep the original effects. But this problem does not exist in the digital future layout.

(4)In regard to printing: similar to visual browsing, the double layer PDF keeps a 100% original visual effects, but because of the limited accuracy of the picture layer, it can't be enlarged too much and printed, which will produce a mosaic blurred. Amended definition TIF images can be printed on large format. Reconstructed PDF support any enlarged font printing with clear and smooth edges, with no distortion and blur, making good print quality than double layer PDF.

(5) On the positioning and retrieval of text: they both support character retrieval and positioning, but the retrieval speed of the double layer PDF is slower than the reconstructed PDF because of the larger file size of double layer PDF

(6) On the storage capacity: the storage capacity of the reconstructed PDF is 1⁄4 to 1⁄6 smaller than double layer PDF. Therefore, opening and network transmission of reconstructed PDF is faster than double layer PDF, and thus is more suitable for web browsing.

(7)On the text error rate: in the theory, the text error rate is not related to the use of double or reconstruction PDF, the error rate only relates to the OCR recognition accuracy and manual correction and so on. In this regard, the difference of double and reconstruction PDF lies on: for double PDF, even there is a text layer typo which is hidden and cannot be seen, it will be reflected in the text retrieval and replication; while for the reconstruction of PDF, if there is an error in text, it can be seen directly. The error rate depends largely on the correction of original scan, text recognition accuracy, the responsibility of proofreading staff, news sense, historical experience and the project management experience of the undertaking company.

(8) On the distribution channels: double PDF is suitable for viewing on the local computer and local area network. Reconstruction of PDF is suitable for viewing on the Internet, mobile phones, tablet PCs, and the large outdoor screen in addition to local computer and local area network,

(9) On the album producing: both technologies are able to meet the individual needs of the album producing.

(10) Costs: the production cost of reconstruction PDF is about 15% ~ 20% higher than the double PDF due to the relatively large production work

Overall, the implementation of the historical newspaper digitization project, if only for protecting and archiving, should scan the historical newspaper layout and create an image database, and then perform character recognition, correction, indexing, storage, and the creation of full-text database and retrieval website; If to further meet the needs of format searching and PDF browsing, double-PDF technology should be adopted; If considering the application of future media terminal (such as Apple's iPhone, iPad tablet PCs), the development of more derivative products, reconstruction of PDF technical solutions should be adopted.

4. Conclusion

The fundamental reason why the vast majority of other historical newspaper databases still use the PDF format and the text operation cannot be implemented lies in the rough original newspaper printing technology the history of the type of information the original printing technology, nonstandard font sizes, and low recognition rate of the historical newspaper resources, thus artificial processing is needed. This means higher human resources and financial costs, and technological breakthroughs are on a broader level exploring.

The digitization of historical newspaper is a project which shows the respect to history, the protection of historical data, and the mining of data value. The library allows the dusty precious newspapers to be presented to the readers in a new look through modern information technology. It reflects the spirit of social responsibility and cultural innovation, co-existence of protection and development, and the librarians' responsibility to make the historical newspapers face the public once more.