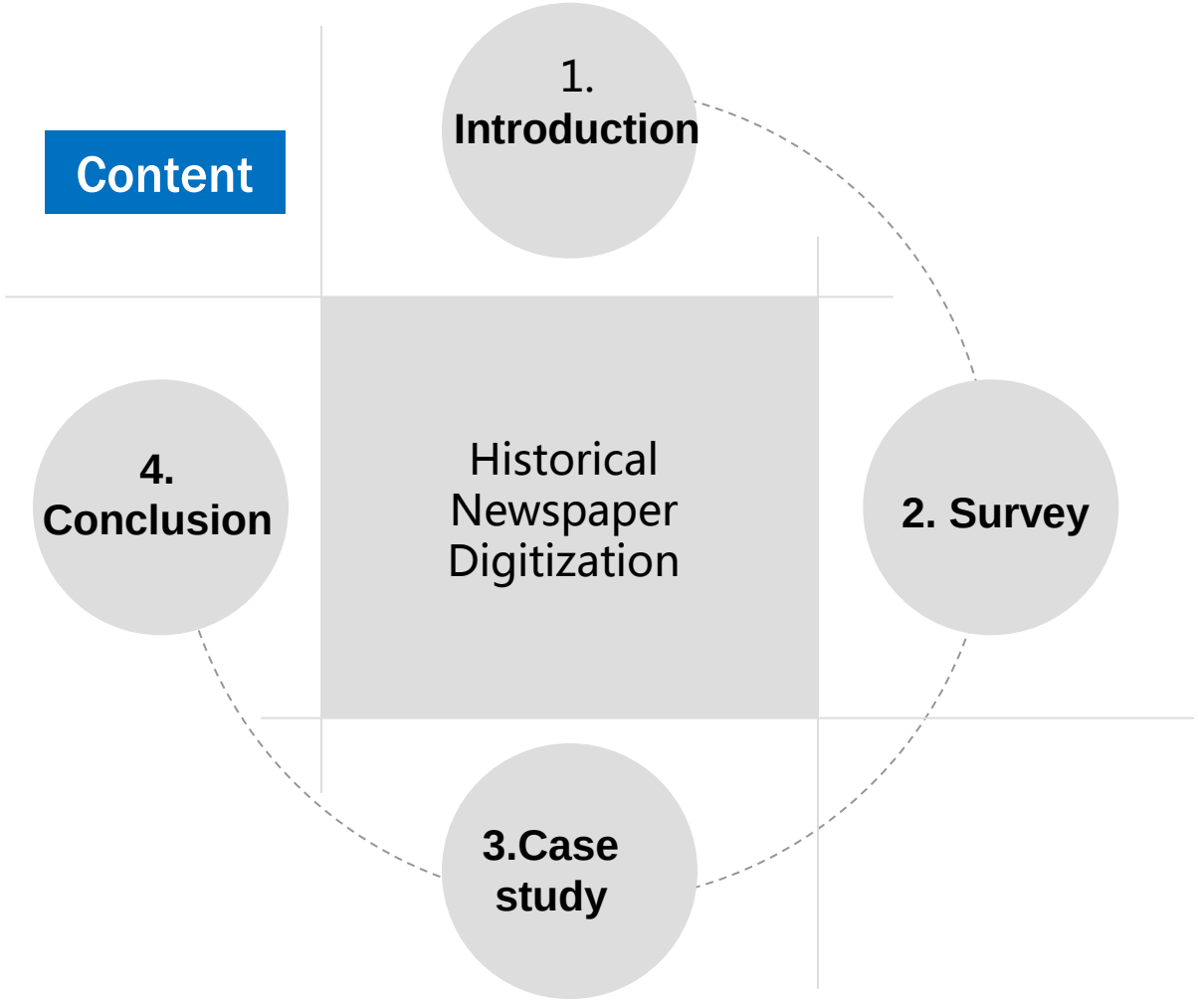
PDF Converter Production of Historical Newspaper Digitization: the picture experience of China's DaChengLaoJiu Database

Reporter: HUANG Weiqun; DING Xiaowen

2014.8.14


Content
1.
Introduction
2. Survey
3.Case
study
4.
Conclusion
Historical
Newspaper
Digitization

1

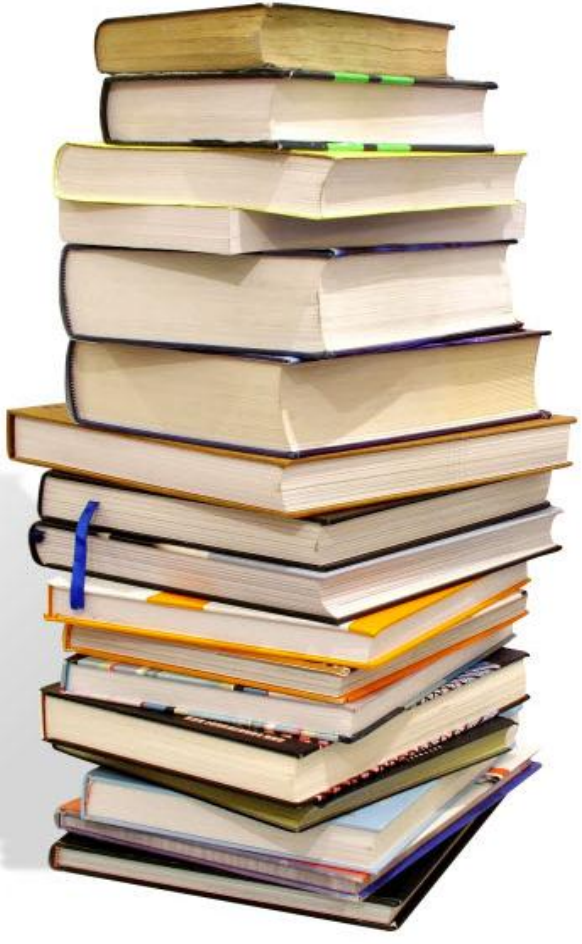
Introduction

Historical newspaper digitization means the display of original historical newspapers, articles and pictures on screen via computer and web technology

2010 China's DaChengLaoJiu
2011 Dazhong Daily historical newspaper digitalization
2012 the digitalization converter production in Beijing Company

2

Survey



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

3

Case study

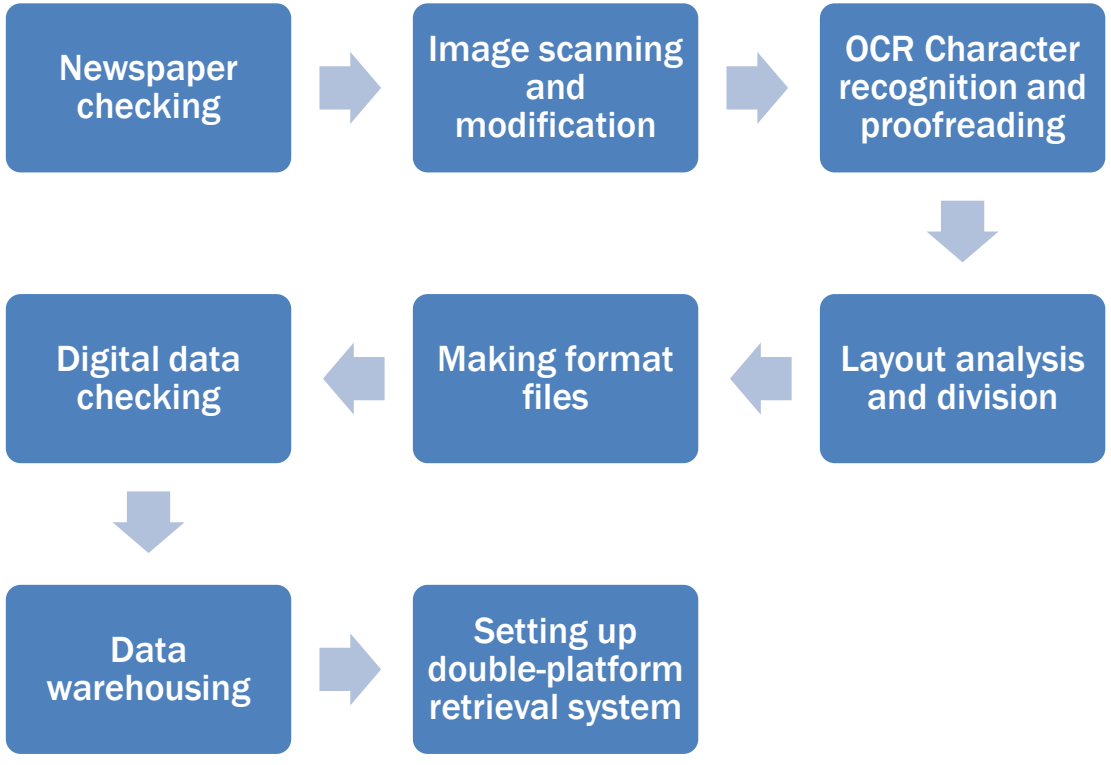
The cases of PDF format files production of Dazhong Daily and Beijing Company

For the early historical newspapers, as there are no corresponding electronic files, so you need to make double layer or refactor the PDF.

Double PDF production :
1.scanning images and processing them into compressed images of appropriate clarity which will be used as the upper image layer of double PDF;
2. rearranging the text according to the original layout structure to form the hidden lower text layer.

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.

Working process



Double PDF ，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).

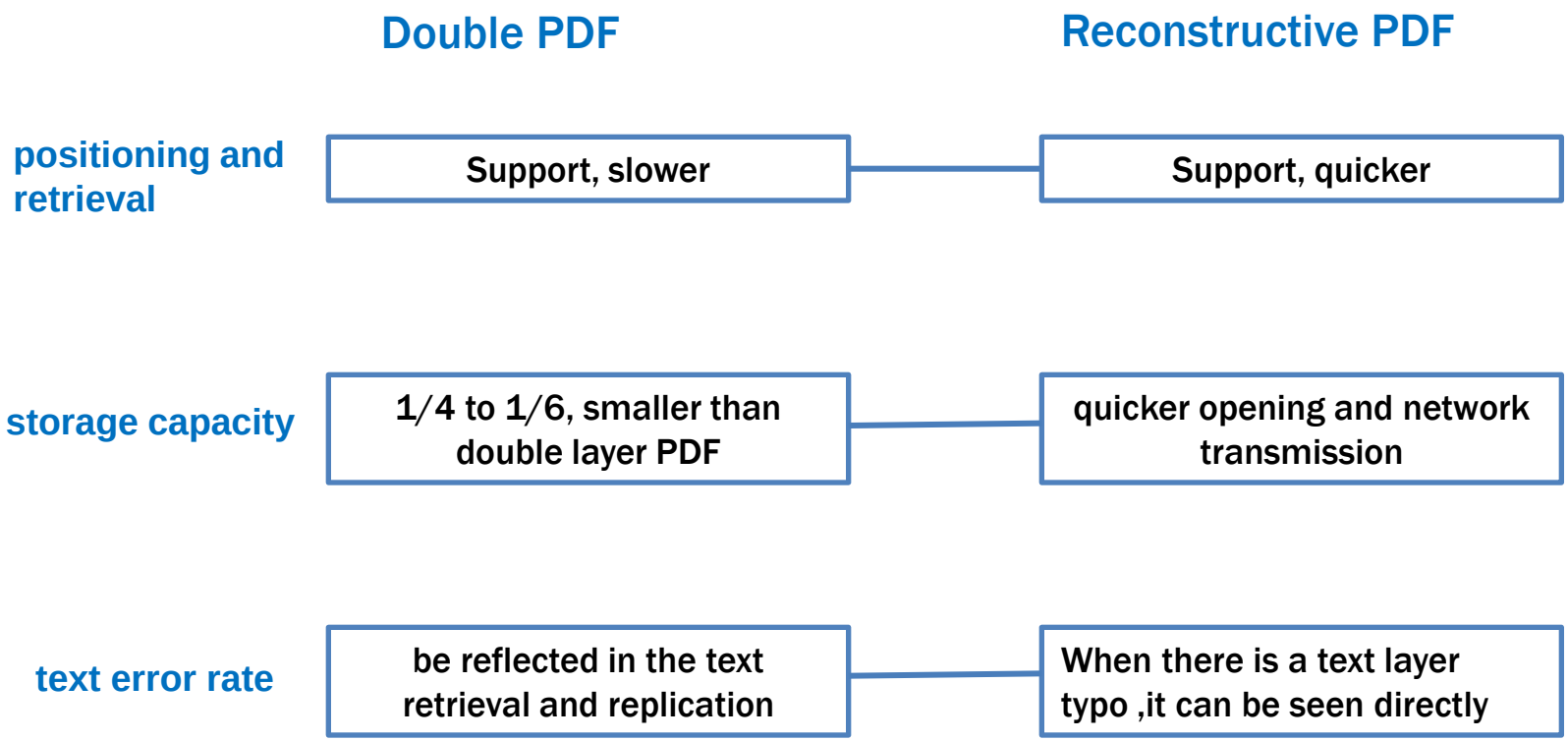
Reconstructive PDF

is a contemporarily popular single image-text structure.

Differences

	Double PDF	Reconstructive PDF
format rearrangement	rearranged according to the original layout	follow the way of today's image-text
Visual browsing	100% maintains scanned layout visual effects; mosaic blur when enlarged	keep the perfect visionary effect; text fonts may be different from the original font.
printing	100% maintains scanned layout visual effects; mosaic blur when enlarged	support any enlarged font printing with clear and smooth edges, with no distortion and blur, good print quality

Differences



Differences

	Double PDF	Reconstructive PDF
distribution channels	suitable for viewing on the local computer and local area network	suitable for viewing on the Internet, mobile phones, tablet PCs, local computer and local area network
album producing	meet the individual needs	meet the individual needs
costs	Cheaper than reconstructive PDF	15% ~ 20% higher than the double PDF due to the relatively large production work

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.

Differences

4

Conclusion

problems

rough original newspaper printing technology the history of the type of information the original printing technology,
nonstandard font sizes
low recognition rate of the historical newspaper resources, thus artificial processing is needed.

higher human resources and financial costs, and technological breakthroughs are on a broader level exploring.

meaning

✓respect to history
✓protection of historical data
✓mining of data value
✓**the spirit of social responsibility and cultural innovation, co-existence of** protection **and development, and the librarians' responsibility**

THANKS *for your time*