(Zhang 2019)

Zhang Wei [张玮]. 2019. *“Minguo baozhi shuzihua yanshou changjian wenti yanjiu——yi guojia tushuguan wei li”* [民国报纸数字化验收常见问题研究——以国家图书馆为例] [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]. *Tushu qingbao yanjiu* [图书情报研究], no. 3: 72–79.

1 Introduction

Acid paper -> urgent need for a new storage method to save the newspapers of the Republic of China, therefore creation of databases based on characteristics, specialities, and commercial use

Sichuan Uni Lib selects by numbers of microfilms used by users, established a platform 《新新新

闻》"New New News" shared platform, searching by Gregorian and Republican calendar, main/subtitles etc

Chongqing Lib started 2012 scanning of Republican newspapers into pdf format

Since 2014 National lib. digitization of newspaper microfilms, based on image scans created article title index

digital publishing companies created large-scale full-text databases, problems such as uneven development of various types of documents and lack of databases for subject types

The digitization of newspapers in the Republic of China has become the primary way to solve newspaper problems

How to improve the efficiency and quality of digitization > Xiao Hong (2017), but the article has no detailed analysis and introduction on quality inspection

2 common problems

digitized products usually include metadata 元数据 and object data对象数据.

Metadata: in MARC format

Object data: TIFF files, publishes jpg, dual-layer pdf, xml

NLC follows standards, uses professional verification software 专业数字化验收软件 developed by Hanwang 汉王

2014-2018 completed 114 newspapers, ca. 130.000 issues, >700.000 pages

Metadata: categories, labels, etc

Item data: file name, image quality etc. coordinates of boxes, content descriptions recorded in XML file

Use of Hanwang software for proofing

Problems: columns, titles, nesting articles, bounding boxes, authors, continued articles, images/image groups

Reasons for problems

1 Personnel lacks experience, problems in coordination of workflow

most of the processing personnel recruited by outsourcing companies外包公司 are junior high school students and high school students大部分是初中生、高中生.

not systematically trained, limited understanding of the tasks

communication issues within the companies, between different “assembly lines”

2 communication between digitisation staff and proofing staff not in time. Changes in the library policies to not get smoothly communicated via company supervisors to digitization staff

3 batch processing leads to stress quantity instead of quality. Repetitive processing enforced from same personnel, but leads to more errors

4 external companies lack experiences and equipment, imperfect results, ocr errors, problems with character recognition are show stoppers

Recommendations:

1 choose between companies, take the best based on previous experiences/tasks

2 calculate error rates and give outcome back if quality does not match

3 processing staff needs to be trained on specifications, in special cases have 1:1 training to get direct communication

4 train proofing personnel, all need to be familiar with the whole process, conduct visits at the external companies to see problems immediately

5 use proof readers according to their best capabilities

G'Translate

Research on Common Problems in the Digital Acceptance of Newspapers in the Republic of China——Taking the National Library as an example  
Zhang Wei  
(National Library, Beijing 100081)

Abstract [Purpose/Meaning] The newspapers of the Republic of China contain a wealth of information about the times, reflecting information about politics, economy, law, culture, military, and education during the period of the Republic of China, and have high research value. This article starts from the perspective of hands-on practice, takes the digital acceptance process as the main line, sorts out the common problems in the digital acceptance of newspapers in the Republic of China, and hopes to provide references for colleagues engaged in digital acceptance in the industry. [Method/Process] From the metadata and object data, it sorts out and summarizes the common problems in the digital acceptance of newspapers from the National Library of China in detail. [Results/Conclusions] The common problems in the digital acceptance of newspapers in the Republic of China are mainly due to the following reasons: digital processing personnel have insufficient experience, uncoordinated coordination between assembly line processing; delayed communication between digital processing personnel and acceptance management; processing People pursue quantity, not quality; outsourcing companies have limited technology and experience. Aiming at the problems and reasons in the digital acceptance of newspapers in the Republic of China, we propose improvement measures and methods: choose the best outsourcing company; if the error rate is higher than a certain threshold, return to the outsourcing company and rework to reduce the problems in the processing of newspapers in the Republic of China from the source; There is a seamless link between technical communication and actual processing personnel; regular training of acceptance personnel to reduce the return rate of digital acceptance errors; according to the experience or expertise of the acceptance personnel, each person’s acceptance tasks are allocated to improve the efficiency of acceptance.  
Key words: newspapers of the Republic of China; digital acceptance; metadata; object data  
Classification Code G255.3  
Cite this article's format Zhang Wei. Research on the Common Problems of the Digital Acceptance of Newspapers in the Republic of China——Taking the National Library as an Example[J]. Library and Information Research,2019,12(3):72-79.

1 Introduction  
The newspapers of the Republic of China contain a wealth of information about the times, reflecting information about politics, economy, law, culture, military, and education during the period of the Republic of China, and have high research value. However, due to the high acidity of newspaper papers in the Republic of China, the current degree of aging and damage of newspapers in the Republic of China is extremely high. There is an urgent need for a new storage method to save the newspapers of the Republic of China. At present, various libraries and publishing houses are actively taking different measures to carry out the Republic of China. The rescue work of newspapers, such as the establishment of a characteristic database of newspapers and periodicals during the Republic of China, a special database of newspapers and periodicals in the Republic of China, a commercial database of newspapers and periodicals in the Republic of China, etc., based on rich, rare, and characteristic newspapers and periodicals in the Republic of China. For academic or commercial use[1]; Sichuan University Library, from the practical perspective of increasing utilization, screens and digitizes the Sichuan newspapers in the Republic of China according to the frequency ranking of the films read by readers; in addition, Sichuan University Library also Independently developed the "New New News" sharing platform, based on the previous data collection and entry, which can support advanced retrieval of the Gregorian calendar date, Republic of China date, newspaper main, subtitle and other fields [2]; Chongqing Library has opened the Republic of China newspaper in March 2012 Digital scanning work, processing the Republic of China newspapers into electronic documents in PDF format that are convenient for readers to retrieve and save [3]. Since 2014, the National Library of China initiated the digitization of newspaper microfilms in the Republic of China [4]. Based on scanned images, it has performed OCR text recognition on article titles [5] to facilitate readers to achieve accurate reading based on title retrieval. As the country attaches great importance to the protection and utilization of the literature of the Republic of China, various digital publishing companies have developed a number of large-scale full-text databases of the Republic of China literature with rich resources and powerful search functions. However, there are problems such as uneven development of various types of documents and lack of databases for subject types. [ 6].  
The digitization of newspapers in the Republic of China has become the primary way to solve newspaper problems. How to improve the efficiency and quality of digitization? **Xiao Hong** analyzes the digital quality inspection issues from the macro perspective (newspaper, format, OCR) [7], but the article has no details on quality inspection Give a detailed analysis and introduction. From the perspective of personal practice, this article takes the digital acceptance process as the main line, lists some common problems found by the National Library in the acceptance of digital products of the Republic of China newspapers, and analyzes the reasons for these problems, and proposes targeted improvement measures and methods. Hope to provide reference for colleagues in the industry who are engaged in digital acceptance work.

2 Types of common problems in the digital acceptance of newspapers in the National Library of the Republic of China  
The digitized products of newspapers in the Republic of China usually include metadata and object data. The digitized metadata of newspapers in the National Library of the Republic of China is recorded in MARC format, which mainly records the characteristics of the literature of the newspaper in the Republic of China, the form of the carrier, and the source of records. The object data adopts TIFF as the storage format and JPG, double-layer PDF and XML as the publishing format, which is a true portrayal of the contents of the newspapers of the Republic of China.  
The digital acceptance of the Republic of China newspaper mainly completes the review of the metadata and object data of the Republic of China newspaper. The National Library of China conducts inspection and acceptance of metadata and object data in accordance with relevant standards, and uses professional digital acceptance **software developed by Hanwang** for acceptance. Since the launch of the Republic of China newspaper digitization project in 2014, as of the end of October 2018, the National Library has completed the acceptance of 114 types of Republic of China newspapers with about 130,000 issues and more than 700,000 editions. Based on actual work, the author summarized the common problems that occurred during the digital acceptance of newspapers in the National Library of China, as shown in Figure 1.

Figure 1 Frequently Asked Questions about the Digital Acceptance of National Library and Republic of China Newspapers

2.1 Frequently Asked Questions about the Acceptance of Newspaper Metadata  
From the perspective of acceptance specifications, the acceptance of the metadata of newspapers in the Republic of China usually checks the legality of labels such as "record identification", "title", "publishing date", "volume", "version", and "version name". Whether the content of the description complies with the rules, such as: the publication date must be 8 digits, and the volume period can only be in digital format. In actual work, the author found that the common problems in the acceptance of the metadata of newspapers in the Republic of China include: illegal character set, missing or redundant labels, and the content of label description does not meet the rules.

(1) The character set is illegal. For example: the label "Author" in Figure 2 is an illegal character set, and the correct identification should be "Image Author".

Figure 2 Illegal character set

(2) The label is missing or redundant. For example: Figure 3 is missing the label "subtitle" and content.

Figure 3 Missing label

(3) The content of label description does not meet the rules. For example: Remarks of the edition in Figure 4: The title is incomplete. Remarks on the revision of the description of the revision of the revision, without the description method; the "Publication Date" in Figure 5: June 15, Republic of China 47, the description is wrong, the description can only be 8 valid figures; "Publication date" in 6: 19580631 does not match the recording method of the date; "Volume" in Figure 7: five hundred forty-five, the description is wrong, the description can only be in digital format.

Figure 4 Version remarks are not in compliance  
Figure 5 Irregular publication date  
Figure 6 Irregular publication date  
Figure 7 Volume irregularities

2.2 Data acceptance issues of newspaper objects in the Republic of China  
From the perspective of acceptance specifications, the acceptance of the object data of the Republic of China newspaper includes checking whether the image storage folder and image file name of the object data are correct, whether the image is clear, whether the brightness is moderate, whether the image has been corrected, and whether the image truly reflects the original content , Whether there are missing versions, wrong versions, missing data, etc. In addition, the key point of acceptance is mainly the xml file, which mainly checks whether each label of the xml conforms to the processing specification, whether the title is correctly marked, and whether the content of the title description is correct, etc. The acceptance **software developed by Hanwang** can faithfully reflect the description information of xml.  
The XML file processed by our library (see Figure 8) covers comprehensive newspaper information (common problems are described in section 2.1), the location coordinates of the article, and the title information of the article. For the acceptance of the title position and the description of the title information, you can intuitively find errors with the aid of acceptance software. Common problems in the acceptance process of the newspaper object data of the Republic of China include: column errors, page errors, picture group errors, etc.

Figure 8 The Republic of China newspaper XML file

(1) Column error  
Columns are usually generalized overviews of newspapers, magazines, and other publications based on text content (namely titles), picture types, etc., such as news, sports, entertainment, economy, etc.; the titles or pictures it leads should all belong to the same column, and the column Should not be attributed to a certain article.  
In the digital processing of newspapers in the Republic of China, all the articles and pictures under a certain column should be marked with column information, and the column should not be included in any article. Common problems include: errors in determining whether it is a column, the scope of the column is incorrectly drawn, and the font recognition of the column is incorrect.

① The certainty of the column. For example: Figure 9 has two columns for labeling, "News" and "Foreign Ministry", but the digitally processed xml file column is missing, and the column is not described. Figure 10 The digitally processed xml file column has a wrong description.

Figure 9 Columns are missing  
Figure 10 Column description error

②The scope of the column is wrong. For example: Figure 11 is wrong to divide the column into the title box; Figure 12 Size column  
The order is wrong, and the column with a large overall scope should be written before the column with a small scope.

Figure 11 Columns entered by mistake  
Figure 12 The order of the size column is wrong

③The text recognition of the column is wrong. The text recognition errors in the column usually include: too many characters or fewer characters, wrong character recognition (variant characters, similar characters, simplified and traditional characters, not recognized according to the original appearance, etc.), more symbols and fewer symbols, etc. For example: Figure 13, the text recognition of the column is incorrect.

Figure 13 Column text recognition error

(2) Mistakes  
The title refers to the independent text content of a certain topic under the framework of a certain column. The title includes the title, author, content of the title, reprint information, and drawings. Title setting requires complete, accurate, and aesthetically pleasing information related to the title. The title of the article is subdivided into quotation, title, subtitle, and subtitle; our library divides the quotation, title, and subtitle according to the size of the font. The largest font size is the title. The title is the quotation before the title, and the subtitle after the title. The title is a small standard. For the author of the title, it is required to remove the attributes of the author, such as nationality, office location, etc., and not to be a collective author, but the description of multiple authors must be clearly defined. Article reprinting refers to the incomplete publication of the current article due to the layout and other reasons, and it is reprinted to other pages, and the reprint information is required when the article is identified, so as to fully record the current article. The accompanying drawings refer to the pictures, tables and other forms involved in the current content, which are displayed in the form of additional images. The accompanying drawings are part of the content.  
Title errors are usually divided into: errors in setting the title or scope of the article, unsightly borders of the title, deterministic errors in the title and unsightly title borders, errors in the determinism of the author and errors in font recognition, errors in reprinting, and drawings The error is as follows.

① The title is incorrectly marked.  
A. The content is missing or redundant. For example: Figure 14, the content of the content is missing the "Chapter VII Judicial" part.

Figure 14 Missing content

B. The nesting problem  
Page nesting, that is, one article is in the content coordinates of another article, and the tag "intercept coordinate" needs to be added under the large article to indicate that there is an independent small article in the large article, and the large and small articles are identified separately. For example: Fig. 15 Small article: "Fortune teller" is nested under the large article, and the large article does not have an internal interception coordinate, which leads to the repeated labeling of the small article: "Fortune teller".

Figure 15 Page nesting

②The border of the article is not beautiful.  
A. The broken line of the title frame. For example: Figure 16, the title and author are both marked out of the article frame, which not only has poor aesthetics, but also gives people the illusion that the purpose of the article is not the same; Figure 17, when the article is marked, there are redundant broken lines, which leads to the article frame. It is cumbersome and complicated, and unsightly.

Figure 16 The border of the article is not beautiful  
Figure 17 The broken line problem of the title frame

B. The text box is compressed. For example: in Figure 18, the text box is pressed at the bottom left.

Figure 18 Text box pressing

③The title is wrong  
A. The order of the quoted question, title, and subtopic is incorrectly identified. For example: Figure 19, the quotation and the subtopic are reversed. After the title (the largest word) is judged by the font size, the quotation before the title and the subtopic after it are judged.

Figure 19 Title order recognition error

B. The subtitle is not recognized. For example: Figure 20, subtitles one, two, and three are not recognized.

Figure 20 Subtitle is not recognized

C. The title box is marked incorrectly. Usually include: title box compression, title box incomplete, title box cross, title box is divided into other content (such as author, content, etc.), for example: Figure 21.

Figure 21 The title tagging error

D. The title word recognition error, the error type is the same as the word recognition part of the column.

④ Author error

A. The author's description is wrong. For example: Figure 22, the attributes of the author: France, should be removed; the description methods (writing, translation) of multiple authors should be identified.

Figure 22 The author's description method is wrong  
Figure 23 Author is not fully recognized

B. Multiple authors are not fully recognized. For example: Figure 23, the author is not fully recognized, the circles in the figure are all authors;

C. Author's font recognition error. The error type is the same as the word recognition part of the column.

⑤ Conversion error  
A. Missing transfers. For example: Figure 24, the end of the article: The fourth edition is continued below, indicating that there is reprint information, and the reprint information should be recorded truthfully.

Figure 24 Missing transfer information

B. Transformation was made by mistake. For example: Fig. 25, the content under the transfer is an independent title. In this case, the transferred title will not be transferred, but will be recognized as an independent title on the page.

Figure 25 The independent content of the converted version

C. Error in linking to version. In the process of digital processing, there may be errors in the transfer version due to omissions or misjudgments. For example, in Figure 26, the version should actually be transferred to the third version, and it was written as the second version.

Figure 26 Linking error of version transfer

D. The drawings are marked incorrectly in the reprint, and the information of the reprinted drawings should be uniformly included in the drawings in which the reprint is located. For example: Figure 27, the reprinted drawings should be classified in the accompanying drawing group.

Figure 27 Errors in the transfer of the drawings

⑥ Pictures are wrong  
A. The picture of the title is mistakenly entered into the picture group. For example: Figure 28, the drawings in the article are in the picture group (see the next section for the introduction of pictures).

Figure 28 The attached picture is mistakenly entered into the picture group

B. In the case of multiple drawings. For example, in Figure 29, the two drawings in the title should be recorded separately and truthfully.

Figure 29 Mistakes in the description of multiple drawings

(3) Picture group error  
The picture group is used to record independent pictures, tables, inscriptions, etc., and it has a parallel relationship with the body group; a complete picture also includes information such as picture label, picture title, picture author, etc.; picture as a complete article, In addition to the above, the error types include the following points.

①The picture has a clear text description and is recognized as a title. For example, as shown in Figure 30, the title of the picture should not be recognized as: [Picture], and should be recorded according to the title in the picture.

Figure 30 Picture title recognition error

②The picture does not have a clear title, self-made title or summary. For example: Figure 31, the picture has no clear title, use the general idea as the title, or use self-made titles [Photo], [Comic], [News Picture], [Advertisement], etc.

Figure 31 Picture without title

3 Causes of problems in the digital acceptance of newspapers in the Republic of China and their improvement methods  
Analyze these common problems in the digital acceptance of newspapers in the Republic of China, mainly due to the following reasons:

(1) The digital processing personnel lack experience, and the coordination between assembly line processing is not coordinated. On the one hand, most of the processing personnel recruited by outsourcing companies are **junior high school students and high school students**. Due to the compact processing tasks, the processing personnel are not systematically trained to start operations, and the processing personnel have limited understanding and operation capabilities of the processing tasks, resulting in some Simple operation makes mistakes; on the other hand, the digital processing of outsourcing companies adopts assembly line operation, and the communication between the assembly lines is lacking, and the first assembly line often fails to operate, and the second assembly line is entered according to their own experience.  
(2) The communication between the digital processing personnel and the acceptance management party is not timely. As the requirements for digital finished products in our library continue to increase, minor changes have often occurred in the digital processing policy of newspapers in the Republic of China. Our library communicated to the relevant person in charge of the outsourcing company. The person in charge caused uncoordinated policy delivery and implementation due to technical understanding or untimely communication. Thus there is a problem.  
(3) Digital processing is carried out in batches, processing personnel pursue quantity rather than quality. Outsourcing companies often assign full-load workloads to processing personnel due to the trend of interest. This leads processing personnel to mechanically perform the same operation for a long time every day in order to complete the assigned tasks. Words, more symbols, less symbols) are ignored, resulting in an error.  
(4) Outsourcing companies have limited technology and experience. Due to the limited technology of outsourcing companies, there are often problems such as slight jaggedness in the labeling of articles, which affects the beauty; outsourcing companies lack processing experience and insufficient OCR character recognition font storage, and they will be unable to start when they encounter some character recognition problems.

For the above reasons, the author believes that improvements can be made from the following aspects:  
(1) Choose an outsourcing company based on the best. When selecting an outsourcing company for bidding, we should refer to the company's performance and processing tasks in previous years for screening, and choose one or two companies with rich experience and strong technical level to be responsible for digital processing tasks.  
(2) Calculate the error rate of the outsourcing company's processing according to different types of errors (such as word errors and labeling errors). If the error rate is higher than a certain threshold, the company will be returned to the company and reworked to limit the outsourcing company’s The processing quality reduces the problems existing in the newspaper processing of the Republic of China from the source.  
(3) Processing personnel will be trained in real time according to the processing specifications communicated by our library, and one-to-one training and learning will be conducted on temporary changes or supplementary specifications for special issues, so as to achieve a seamless link from technical communication to actual processing personnel.  
(4) Acceptance personnel are regularly trained, everyone is familiar with the acceptance process, and through on-site visits to the outsourcing company's processing process, the outsourcing company's digital processing problems can be discovered in time and the return rate of digital acceptance errors can be reduced.  
(5) According to the experience or expertise of the inspectors, assign each person's acceptance tasks to improve the efficiency of acceptance.

References:  
[1] Lu Yijun. Research on the Collection Construction of Newspaper and Periodical Resources in Chinese University Libraries during the Republic of my country[J]. Journal of Library and Information Sciences of the University, 2017,35(4): 39-45.  
[2] Jin Caihong. Digitization of newspaper microfilms in Sichuan during the Republic of China: Taking Sichuan University Library as an example [J]. Journal of the Library Science in Sichuan, 2016(3):17-19.  
[3] Chen Guixiang. On the digital construction of newspapers in the Republic of China——Taking Chongqing Library as an example[J]. Science and Technology Information Development and Economy, 2013, 23(4): 27-29.  
[4] Xiao Hong, Wu Ming, Zeng Yan. A Probe into the Digitization and Service of Microfilms in the Republic of China Newspapers: Taking the National Library as an Example [J]. Library Science, 2015(10): 89-92.  
[5] Xiao Hong. The main problems and solutions in the practice of newspaper digitization in the Republic of China[J]. Research in Library Science, 2017(4): 22-37.  
[6] Duan Xiaolin. Research on the Development Status of the Literature Database of the Republic of China [J]. Research in Library Science, 2016(20): 42-45.  
[7] Xiao Hong, Huai Yan. Analysis of quality inspection problems in the practice of newspaper digitization in the Republic of China[J]. Library Science Research, 2017(7): 61-78,87.

About the author Zhang Wei, female, born in 1987, master's degree, librarian, research direction is document digital processing and processing, and published several papers.